From Pixels to People: Recovering Location, Shape and Pose of Humans in Images

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
Mohamed Omran
M.Sc. Visual Computing

Saarbrücken
2021

Date of Colloquium	15th of December, 2021
Dean of the Faculty	Univ.-Prof. Dr. Thomas Schuster Saarland University, Germany

Examination Committee

Chair	Prof. Dr. Vera Demberg
Reviewer, Advisor	Prof. Dr. Bernt Schiele
Reviewer	Prof. Dr. Jürgen Gall
Academic Assistant	Jan-Eric Lenssen

Abstract

Humans are at the centre of a significant amount of research in computer vision. Endowing machines with the ability to perceive people from visual data is an immense scientific challenge with a high degree of direct practical relevance. Success in automatic perception can be measured at different levels of abstraction, and this will depend on which intelligent behaviour we are trying to replicate: the ability to localise persons in an image or in the environment, understanding how persons are moving at the skeleton and at the surface level, interpreting their interactions with the environment including with other people, and perhaps even anticipating future actions. In this thesis we tackle different sub-problems of the broad research area referred to as "looking at people", aiming to perceive humans in images at different levels of granularity.

We start with bounding box-level pedestrian detection: We present a retrospective analysis of methods published in the decade preceding our work, identifying various strands of research that have advanced the state of the art. With quantitative experiments, we demonstrate the critical role of developing better feature representations and having the right training distribution. We then contribute two methods based on the insights derived from our analysis: one that combines the strongest aspects of past detectors and another that focuses purely on learning representations. The latter method outperforms more complicated approaches, especially those based on handcrafted features. We conclude our work on pedestrian detection with a forward-looking analysis that maps out potential avenues for future research.

We then turn to pixel-level methods: Perceiving humans requires us to both separate them precisely from the background and identify their surroundings. To this end, we introduce *Cityscapes*, a large-scale dataset for street scene understanding. This has since established itself as a go-to benchmark for segmentation and detection. We additionally develop methods that relax the requirement for expensive pixel-level annotations, focusing on the task of boundary detection, i.e. identifying the outlines of relevant objects and surfaces. Next, we make the jump from pixels to 3D surfaces, from localising and labelling to fine-grained spatial understanding. We contribute a method for recovering 3D human shape and pose, which marries the advantages of learning-based and model-based approaches.

We conclude the thesis with a detailed discussion of benchmarking practices in computer vision. Among other things, we argue that the design of future datasets should be driven by the general goal of combinatorial robustness besides task-specific considerations.

Zusammenfassung

Der Mensch steht im Zentrum vieler Forschungsanstrengungen im Bereich des maschinellen Sehens. Es ist eine immense wissenschaftliche Herausforderung mit hohem unmittelbarem Praxisbezug, Maschinen mit der Fähigkeit auszustatten, Menschen auf der Grundlage von visuellen Daten wahrzunehmen. Die automatische Wahrnehmung kann auf verschiedenen Abstraktionsebenen erfolgen. Dies hängt davon ab, welches intelligente Verhalten wir nachbilden wollen: die Fähigkeit, Personen auf der Bildfläche oder im 3D-Raum zu lokalisieren, die Bewegungen von Körperteilen und Körperoberflächen zu erfassen, Interaktionen einer Person mit ihrer Umgebung einschließlich mit anderen Menschen zu deuten, und vielleicht sogar zukünftige Handlungen zu antizipieren. In dieser Arbeit beschäftigen wir uns mit verschiedenen Teilproblemen die dem breiten Forschungsgebiet "Betrachten von Menschen" gehören.

Beginnend mit der Fußgängererkennung präsentieren wir eine Analyse von Methoden, die im Jahrzehnt vor unserem Ausgangspunkt veröffentlicht wurden, und identifizieren dabei verschiedene Forschungsstränge, die den Stand der Technik vorangetrieben haben. Unsere quantitativen Experimente zeigen die entscheidende Rolle sowohl der Entwicklung besserer Bildmerkmale als auch der Trainingsdatenverteilung. Anschließend tragen wir zwei Methoden bei, die auf den Erkenntnissen unserer Analyse basieren: eine Methode, die die stärksten Aspekte vergangener Detektoren kombiniert, eine andere, die sich im Wesentlichen auf das Lernen von Bildmerkmalen konzentriert. Letztere übertrifft kompliziertere Methoden, insbesondere solche, die auf handgefertigten Bildmerkmalen basieren. Wir schließen unsere Arbeit zur Fußgängererkennung mit einer vorausschauenden Analyse ab, die mögliche Wege für die zukünftige Forschung aufzeigt.

Anschließend wenden wir uns Methoden zu, die Entscheidungen auf Pixelebene betreffen. Um Menschen wahrzunehmen, müssen wir diese sowohl praezise vom Hintergrund trennen als auch ihre Umgebung verstehen. Zu diesem Zweck führen wir *Cityscapes* ein, einen umfangreichen Datensatz zum Verständnis von Straßenszenen. Dieser hat sich seitdem als Standardbenchmark für Segmentierung und Erkennung etabliert. Darüber hinaus entwickeln wir Methoden, die die Notwendigkeit teurer Annotationen auf Pixelebene reduzieren. Wir konzentrieren uns hierbei auf die Aufgabe der Umgrenzungserkennung, d. h. das Erkennen der Umrisse relevanter Objekte und Oberflächen.

Als nächstes machen wir den Sprung von Pixeln zu 3D-Oberflächen, vom Lokalisieren und Beschriften zum präzisen räumlichen Verständnis. Wir tragen eine Methode zur Schätzung der 3D-Körperoberfläche sowie der 3D-Körperpose bei, die die Vorteile von lernbasierten und modellbasierten Ansätzen vereint.

Wir schließen die Arbeit mit einer ausführlichen Diskussion von Evaluationspraktiken im maschinellen Sehen ab. Unter anderem argumentieren wir, dass der Entwurf zukün-

ftiger Datensätze neben aufgabenspezifischen Überlegungen vom allgemeinen Ziel der kombinatorischen Robustheit bestimmt werden sollte.

Acknowledgements

This thesis took much more time to produce than it should have, and an exhaustive list of people whom I am grateful to is accordingly very long and would fill several pages.

I am grateful to everyone at the Max Planck Institute for Informatics for creating a stimulating and supportive environment. I am especially grateful to Bernt for giving me the opportunity to work here, and for the patience and mentoring he provided especially in the difficult final stretch. I am also very grateful to Connie who kept things running very smoothly on the adminstrative side as well as to the IST and Haustechnik teams who did the same on the technical side. I am also extremely grateful to the IMPRS-CS team headed by Jennifer Gerling for the support during my masters studies.

I would also like to thank all the many collaborators whose valuable work made this thesis possible including in no particular order: Rodrigo Benenson, Jan Hosang, Anna Khoreva, Shanshan Zhang, Marius Cordts, Stefan Roth, Markus Enzweiler, Sebastian Ramos, Timo Rehfeld, Uwe Franke, Christoph Lassner, Peter Gehler, Gerard Pons-Moll, Evgeny Levinkov, Eldar Insafutdinov, Bjoern Andres, Mario Fritz, and of course Bernt Schiele. I also owe much gratitude to early teachers and advisors who helped spark and further my interest in computer vision including Heiko Neumann, Dariu Gavrila, Markus Enzweiler, Christoph Keller and Joachim Weickert.

Countless friends and colleagues made my stay in Saarbruecken memorable and I trust they know how much their company and cameraderie meant and still means to me.

Finally I would like to thank my family who have been unwavering in their support and without whom I would have not been able to finally write these words: Mum, Dad, Ahmed, and Kata, no words could do you justice.

Contents

Introduction

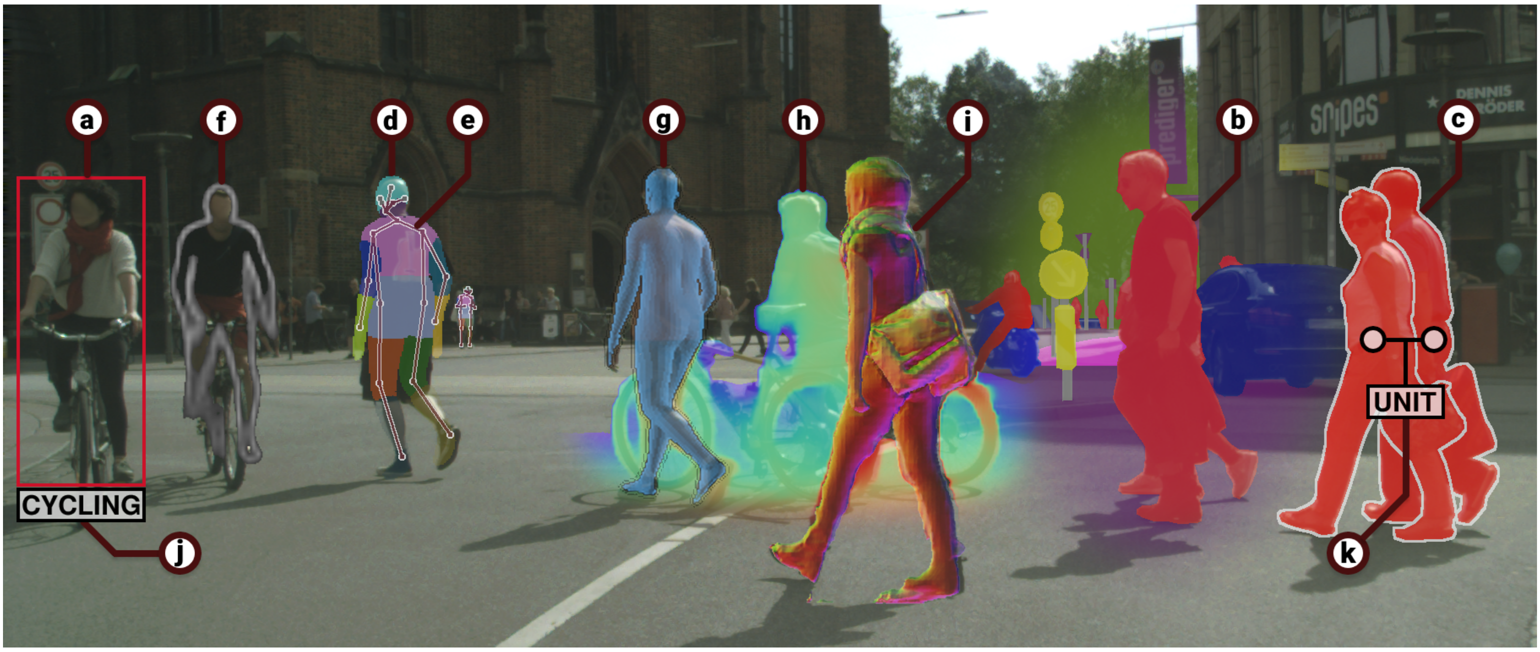


Figure 1.1: What often distinguishes many tasks in human-centric computer vision is the output abstraction extracted from the image. This includes (a) bounding boxes, (b) per-class collections of pixels, (c) per-instance masks — sometimes with (d) fine-grained semantic labels, (e) skeletons both 2D and 3D, (f) 2D contours, (g-i) a variety of 3D surface representations, as well as high-level semantic outputs such as (j) activities or (k) social relations and signals. (See main text for explanations and references.)

1.1 Looking at People

Humans are the objects of study in a significant amount of computer vision research. This is unsurprising, as it would be immensely useful to replicate in machines our own ability to perceive and understand humans from visual data. While the goal of machines with human-like intelligence remains distant and ever-elusive, human-centric computer vision has already delivered advances that are having a substantial real-world impact across a variety of domains, including transportation, human-computer interaction, and the creative arts to name just a few.

To endow intelligent agents, such as autonomous vehicles, with the ability to navigate physical environments safely, they need to detect nearby people. Intelligent interfaces that can respond to basic gestures must go beyond merely localising people. They additionally need to explicitly identify and track individual body parts. Cutting-edge applications in animation, virtual reality, and telepresence are powered by methods that retrieve precise information about the three-dimensional surface of the human body.

From this brief list of examples, one can see that different applications target different output abstractions. In fact, what distinguishes many sub-areas of human-centric computer vision is the final representation that is extracted from the image. See Fig. 1.1[1] for a set of examples. This thesis addresses several tasks broadly related to this domain, otherwise referred to sometimes as "looking at people" (Pentland, 2000; Gavrila, 2007). In the next part of this introduction, we thus give an overview of the sub-areas relevant to this thesis.

1.1.1 Overview

One fundamental task in computer vision is object detection: exhaustively identifying sub-areas of an image that each contain a single object — in our case a person. The output here is conventionally the axis-aligned bounding box (Fig. 1.1a). One variant of this task is pedestrian detection, where the goal is to localise people in street scenes. In Chapters 4 to 6 we address this task.

Pedestrian detection — or more broadly people detection — is a fundamental task as it is the base component for many methods that extract spatially or semantically fine-grained information from images. These methods either explicitly identify image sub-regions that require further processing as an initial step, or assume that persons of interest have been pre-localised.

Object detection only leads to a coarse localisation in the image plane, but it is often useful or even necessary to extract more spatially fine-grained outputs, namely by labelling individual pixels. Many tasks in computer vision in fact can be cast as pixel labelling problems even if this is not immediately apparent, including bounding box-based detection.

Semantic segmentation (Fig. 1.1b) involves assigning a semantic category label to each pixel from a pre-defined set. The set of categories under consideration is application- or dataset-dependent. In a street scene setting, this set can include pedestrians, vehicles, and various kinds of surfaces and structures. This information can provide useful context for localising people. Other useful categories for perceiving and understanding humans can include anatomical parts and clothing (Fig. 1.1d).

[1] Fig. 1.1: (b-c) are ground truth annotations from *Cityscapes* (Chapter 7) as is the image itself, (d-g) were generated by an unpublished follow-up to the method in Chapter 9, (h) was generated by the method of (Li *et al.*, 2019), and (i) is the output from *PiFU-HD* (Saito *et al.*, 2020) with some manual intervention to suppress the background.

When pixels are not only assigned category labels but also instance labels — or in other words, when pixels are grouped if they belong to the same object — we speak of instance segmentation (Fig. 1.1c). Sometimes this task is addressed together with fine-grained semantic labelling and referred to as instance-level human parsing. Grouping pixels into larger units is challenging, and many instance segmentation methods take a top-down or instance-first approach. First they localise objects at the bounding box level, then solve a binary segmentation problem for each identified instance. Bottom-up methods on the other hand, tackle the grouping problem globally with pixel-level information. In Chapter 7, we present a benchmark for pixel-level and instance-level segmentation: *Cityscapes*, which has become the de facto standard in this domain.

Boundaries are important cues for recognition as they can indicate the transition from one object or surface to the next. However, boundaries are hard to define since their perceptual relevance changes depending on the objects we're interested in. One variant of this task is semantic boundary detection. Similar to semantic segmentation, this task focuses on a limited set of categories, but the goal is to only label the outlines of objects and identify the class (Fig. 1.1f). In Chapter 8, we address different variants of boundary detection while relaxing the requirement for hand-annotated ground truth.

Another key problem relevant to our work is 2D pose estimation (Fig. 1.1e). Here, the human body is represented as a sparse set of keypoints corresponding to different locations either inside or on the surface of the human body, e.g. eyes, hands and knees. The task is then to localise the 2D projections of these points on the image.

In the above, we described tasks where the output is strictly two-dimensional. Humans inhabit three-dimensional space and to perceive and reason about them in 3D, we correspondingly need representations that go beyond the image plane. A wide-variety of 3D representations exist in the literature: 3D bounding boxes, 3D skeletons, pixel-wise depth maps that can only capture the depth of visible surfaces, or others that can capture more complete 3D structure, including surfaces not visible to the camera. These include voxels, meshes, and implicit surface functions (Fig. 1.1g-i).

Such representations in principle can represent any surface and don't make explicit assumptions on the structure of the world or the objects occupying it. For humans however, we can exploit prior knowledge, namely structural regularities of the body. Parametric mesh models of the human body allow us to efficiently represent the surface in detail with a smaller number of parameters that separately capture shape and pose variation. In Chapter 9, we present a method for predicting 3D human shape and pose which embeds such a model into a neural network. There, we also demonstrate how reasoning in 3D can benefit from different 2D representations.

1.2 Outline and Contributions

This thesis consists of three parts: The first focuses on pedestrian detection (Chapters 4 to 6). The second part looks at pixel-level classification tasks (Chapters 7 and 8). In the third part, we discuss a higher-level recognition task, namely predicting human body shape and pose (Chapter 9).

Chapter 2: Related Work: Pedestrian Detection. This chapter introduces the task of pedestrian detection, presenting the relevant evaluation metrics and datasets. We then describe common approaches to generic detection, and provide an overview of recent work in area of detecting pedestrians specifically.

Chapter 3: Related Work: 3D Human Shape and Pose Recovery. Here, we discuss the task of 3D human shape and pose estimation. Representing humans in 3D and acquiring the corresponding ground truth is a challenge, so we spend time on describing relevant efforts before reviewing methods.

Chapter 4: Lessons from a Decade of Pedestrian Detection: 2004-2014. In this chapter, we present a survey and quantitative analysis of pedestrian detection that covers methods published between 2004 and 2014 — complementary to the discussion in Chapter 2. We analyse several families of methods and identify those that have led to consistent performance improvements. We combine representative methods from the latter into a single approach and achieve top performance. One key takeaway from this chapter is the overwhelming importance of image representations, as well as using the right data.

Chapter 5: Deep Learning for Pedestrian Detection. Motivated by the critical importance of features for pedestrian detection identified in the previous chapter, we look into deep learning (DL) for pedestrian detection. We show how plain convolutional neural networks (CNNs) can be applied to the task successfully in combination with a strong traditional detector as a proposal method. We demonstrate significant improvements over prior DL-based approaches without resorting to problem-specific modelling.

Chapter 6: Towards a Human Baseline for Pedestrian Detection. This chapter provides a forward-looking analysis complementary to the retrospective analysis of Chapter 4. We establishes a human baseline as an upper bound for pedestrian detection performance on one popular benchmark. This involves a novel annotation protocol for pedestrians that results in more accurate training annotations allowing us to measure the impact of label noise on pedestrian detectors. Furthermore, we analyse the failure modes of state-of-the-art detectors in detail, addressing some of them and suggesting directions for future research.

Chapter 7: The Cityscapes Dataset for Semantic Urban Scene Understanding. In this chapter we present a large-scale street scene understanding dataset: *Cityscapes*. We provide detailed annotations for pixel-level and instance-level segmentation and conduct an in-depth quantitative analysis of the dataset characteristics. We additionally evaluate several baseline methods as well as state-of-the-art approaches.

Chapter 8: Weakly-Supervised Boundary Detection. Here, we look at another pixel-wise classification task, namely object boundary detection. Obtaining annotations for this task is particularly arduous. We thus experiment with different methods that relax the requirement for large amounts of ground truth. The proposed weakly-supervised techniques achieve strong performance compared to both competing weakly- as well as fully-supervised methods on different variants of the boundary detection task.

Chapter 9: Neural Body Fitting: 3D Human Shape and Pose Recovery. We then move on to 3D human shape and pose estimation. Here, we incorporate a statistical body model into a CNN, thus marrying the benefits of direct prediction and model-based approaches. We show that high performance and data-efficient training can also be achieved by breaking the problem down into body part segmentation, followed by a 2D-to-3D lifting step. The presence of the model also allows us to supervise our prediction network with 2D data.

Chapter 10: Conclusions and Future Directions. We summarise the conclusions of this thesis and present a wide-ranging discussion of possible future work. In particular, we take stock of current benchmarking practices in computer vision and suggest promising directions for improving both evaluation and dataset design. With regards to the latter, we suggest the pursuit of combinatorial robustness as a guiding principle besides task-specific considerations. We conclude by discussing the need for the need for more dynamic models that incorporate recurrence and feedback.

2 Related Work: Pedestrian Detection

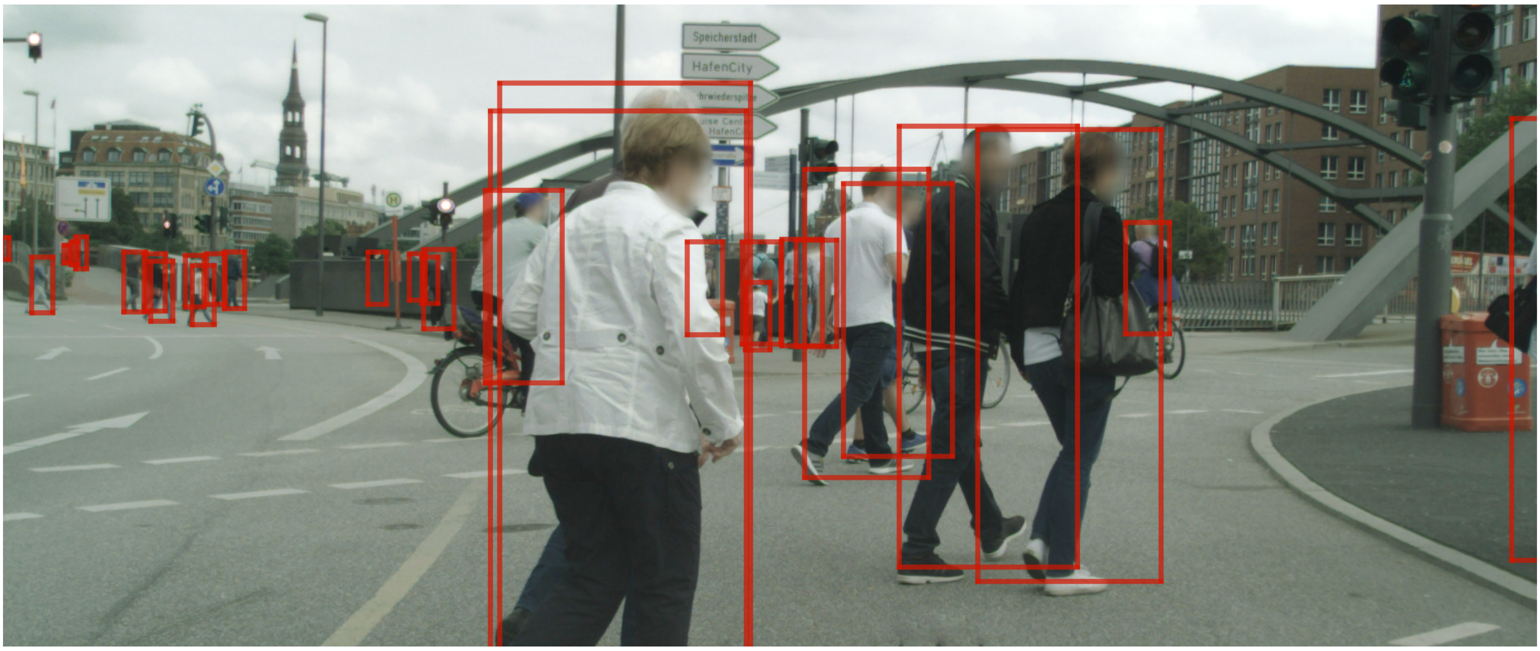


Figure 2.1: Pedestrian detection is a highly challenging task, especially in crowded urban scenes where methods have to contend with strong occlusions and scale variation. (image from *Cityscapes* (Chapter 7), annotations from *CityPersons* (Zhang *et al.*, 2017b))

Pedestrian detection is a canonical computer vision task that continues to reliably attract attention from researchers, with new datasets being released on a yearly basis (Hwang *et al.*, 2015; González *et al.*, 2016; Zhang *et al.*, 2017b; Neumann *et al.*, 2018; Braun *et al.*, 2019; Zhang *et al.*, 2020b). It endures as a problem that is tackled separately from general object detection due to a unique set of challenges. Besides the high variability of pedestrian appearance, other challenges include cluttered environments, difficult recording conditions and high scale variation.

Challenges

Urban environments are often cluttered (Fig. 3.1), containing a variety of objects both static and in motion. These objects can confound detectors as they are sometimes pedestrian-like in appearance, and can also severely occlude pedestrians. A person might for example suddenly appear on the street from behind a parked car or some other obstacle, and thus dealing with occlusions is highly important from a safety perspective. While person-to-object occlusions are challenging enough, person-to-person occlusions are arguably even more challenging, given the difficulty of not just having to recognise occluded persons, but also having to disambiguate different persons when in close proximity to one another.

Images are typically recorded from a vehicle-mounted — and hence moving — camera. This results in artifacts such as motion blur, but also makes it necessary to detect distant pedestrians as one approaches them at speed. Urban computer vision datasets correspondingly exhibit much more scale variation than more traditional object detection datasets (Chapter 7). Dealing with this scale variation has been the subject of much research in this domain. Getting computer vision systems to work in outdoor scenes also means having to contend with a variety of challenging weather and lighting conditions. Keeping pedestrians safe does not just involve detecting them but also tracking and even anticipating their motion (e.g. Rasouli *et al.* (2017)). While these are interesting related problems where detection plays an integral part — and can even benefit from the temporal reasoning involved —, these are out of the scope of this thesis as we focus on the single-frame detection problem.

Summary

In this chapter we survey recent progress in the domain of pedestrian detection. We start with a definition of the task as it is addressed today (Sec. 2.1) and describe the established evaluation metrics and datasets (Sec. 2.2). Contemporary approaches are almost exclusively based on end-to-end neural networks (NN) which integrate search, representation learning, classification and localisation into a jointly-optimised pipeline. In recent years these have gradually displaced the previously dominant approaches that were based on hand-crafted features. In Chapter 4 we present a detailed analysis that mostly covers this older class of methods from the years 2004-2014, and in Chapter 5 an early NN-based approach to pedestrian detection. In Chapter 6, we examine failure modes for both classes of methods. To complement this work, in Sec. 2.3 we first describe different approaches to detection with an emphasis on modern end-to-end detectors. We then trace out the transition from detectors fully reliant on hand-crafted features, through to mixed pipelines combining both classical detectors and NN-based classifiers, up until the most recent crop of end-to-end detectors.

2.1 Task Definition

Given an image, the task is to produce a set of axis-aligned bounding boxes (detections) that can be matched one-to-one with another set of bounding boxes (ground truth).

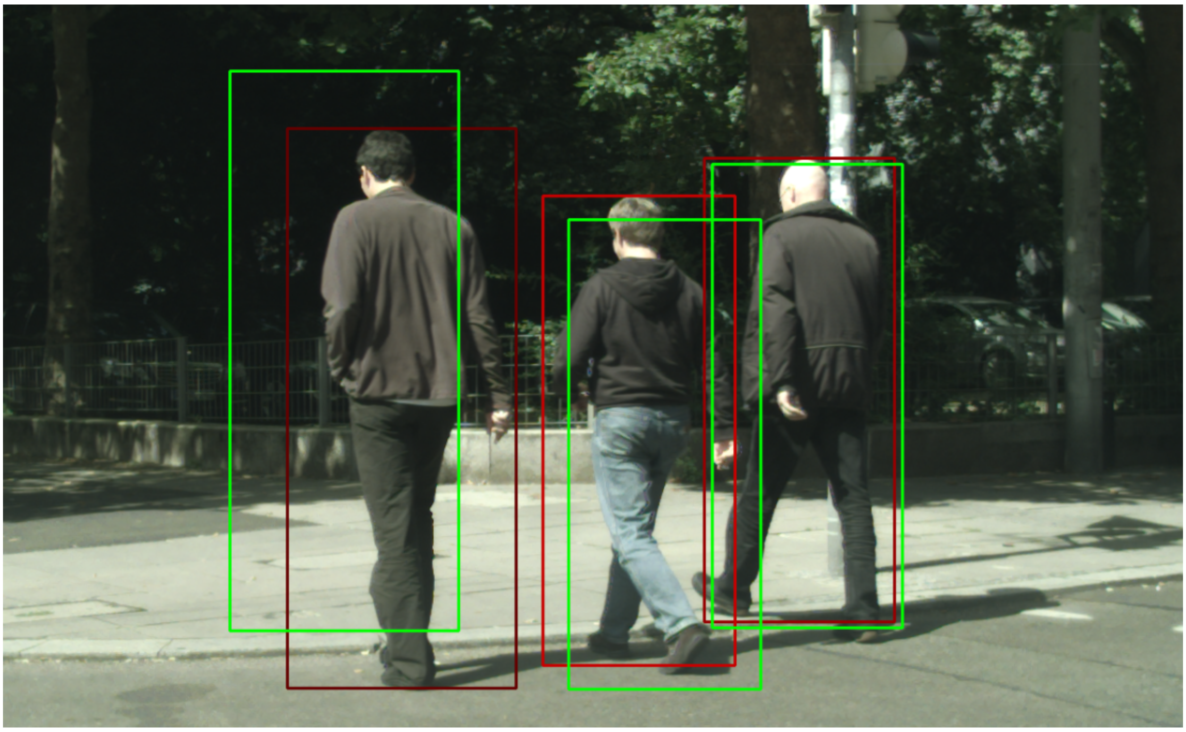


Figure 2.2: Illustration of the pedestrian detection task. The ground truth bounding boxes are in red and detections in green. The latter are all true positives according to the standard evaluation metric, with respective intersection over union (IoU) values of 0.5, 0.7 and 0.9 from left to right. This is meant to demonstrate the looseness of the target evaluation metric.

Each bounding box is associated with a single pedestrian, tightly enclosing it. This is illustrated in Fig. 2.2, in which ground truth boxes are in red and detections in green.

Detectors will typically produce an overcomplete set of detections of which multiple hypotheses can be matched to a single pedestrian based on spatial overlap alone. Thus to evaluate the output of a detector, one proceeds as follows: Hypotheses are ranked according to their confidence score, normally produced by the detector together with the bounding box coordinates. Hypotheses with a confidence score below a threshold c are discarded. Those remaining are then in sequence either matched to an unassigned ground truth box or considered to be false positives if no possible match is found.

A match between detection BB_{dt} and ground truth box BB_{gt} is considered successful if the area of their overlap exceeds a specific threshold. This is conventionally half of the joint area of both boxes, a measure popularised by the *PASCAL VOC* benchmark (Everingham *et al.*, 2015). This is expressed as:

$$\text{IoU} \doteq \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} \tag{2.1}$$

where IoU stands for “intersection over union”.

The successfully matched ground truth box is then removed from consideration and matching proceeds in a similar fashion for the rest of the boxes. For a given confidence threshold c, successfully matched detections are true positives ($TP(c)$), unmatched detections are false positives ($FP(c)$) and unmatched ground truth boxes are false negatives ($FN(c)$).

In pedestrian detection, the selection of an appropriate confidence threshold c conventionally involves a trade-off between the fraction of missed pedestrians ($MR(c)$) and the number of false positives per image ($FPPI(c)$). This choice is dictated by safety considerations in the automotive domain (Dollár *et al.*, 2012b). We wish to reduce the number of false alarms regardless of how many objects are present per frame, since every detection might require a change to the vehicle's path.

The miss rate is given by:

$$MR(c) = \frac{FN(c)}{(TP(c) + FN(c))} \tag{2.2}$$

To visualise the trade-off, we simply plot the miss rate against the number of false positives per image as we vary the threshold c. Performance is then summarised using the log-average miss rate (lower values are better) (Dollár *et al.*, 2012b), which is computed by averaging MR at nine values of $FPPI$, evenly distributed in log-space:

$$laMR = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} MR(c), \tag{2.3}$$

$$\mathcal{C} = \{c | FPPI(c) \in \{10^0, 10^{-.25}, 10^{-0.5}, ..., 10^{-2}\}\} \tag{2.4}$$

The matching process is normally complicated by the availability of ground truth boxes not considered for evaluation for one of two reasons: either (i) because these were explicitly marked as "ignore" regions during annotation, or (ii) they fall outside size or occlusion ranges pre-selected for evaluation. Recent benchmarks all define a so-called "Reasonable" setting, that includes boxes above $40 - 50$ pixels in height and with only a small degree of occlusion, e.g. up to 35%. Harder settings involve smaller and more highly-occluded pedestrians, but performance under the "Reasonable" setting is how methods are often ranked against one another.

This evaluation approach was established as the de facto standard with the publication of the *Caltech Pedestrian Dataset* (Dollár *et al.*, 2009b), further refined in Dollár *et al.* (2012b) and has hardly been modified since. Most subsequently released datasets, such as the *KAIST Multispectral Pedestrian Dataset* (Hwang *et al.*, 2015), *CityPersons* (Zhang *et al.*, 2017b), *NightOwls* (Neumann *et al.*, 2018) and *EuroCity Persons* (Braun *et al.*, 2019), have largely adopted these conventions. Minor modifications include adjusting the size and occlusion ranges for different evaluation settings.

Previous benchmarks such as *INRIA* (Dalal and Triggs, 2005), rather than measuring false positives per image, measured them per window (FPPW). The test set would consist of pedestrian cutouts and negative windows sampled from pedestrian-free images.

The advantage of this measure is that it decouples classification performance from other aspects of detection such as search and non-maximum suppression. The disadvantage is that classification performance is not a good predictor of detection performance (Dollár *et al.*, 2009a) as there are interactions between the different components of a detector. This measure would for example not distinguish between (i) a detector that produces highly localised positive responses when centered on pedestrians, and (ii) another which identifies pedestrians correctly, but produces more diffuse responses causing more difficulties during the non-maximum suppression stage, which removes redundant detections.

This also differs from the evaluation metric used for generic object detection. There, precision is plotted against recall while varying the confidence threshold c. Precision tells us what percentage of positive detections actually belong to the positive class:

$$Pr(c) = \frac{TP(c)}{(TP(c) + FP(c))} \tag{2.5}$$

Recall measures the percentage of positive instances that have been correctly detected:

$$Rec(c) = \frac{TP(c)}{(TP(c) + FN(c))} \tag{2.6}$$

Note that $Rec(c) = 1 - MR(c)$. Older benchmarks such as *PASCAL VOC* (Everingham *et al.*, 2015) summarise performance using average precision (AP), which is computed as the area under the precision-recall (PR) curve — and thus sometimes alternatively referred to as AUC (for area under the curve). The correctness criterion for a detection remains the same as above: $IoU > 0.5$ with an undetected ground truth bounding box, but in contrast to laMR, higher values are better. The *KITTI* benchmark (Geiger *et al.*, 2012) adopts this metric in contrast to most other pedestrian benchmarks. Newer benchmarks such as *MSCOCO* (Lin *et al.*, 2014) report mean average precision (mAP), for which AP is averaged for a range of different IoU thresholds from 0.5 to 0.95.

2.2 Datasets and Benchmarks

Currently, the most widely used benchmarks for pedestrian detection are the *Caltech Pedestrian Dataset* (Dollár *et al.*, 2012b) and *CityPersons* (Zhang *et al.*, 2017b).

The *Caltech Pedestrian Dataset* (*Caltech* or sometimes *Caltech-USA*) (Dollár *et al.*, 2012b) was the first large-scale pedestrian dataset. It consists of 10 hours of 30Hz video ($640px \times 480px$) recorded from a car driving through the Los Angeles metropolitan area. Manual annotations are provided for keyframes (1fps) and annotations for the same person are linked. This allows for automatic annotation of the remaining frames via interpolation, resulting in a total of 350 000 bounding boxes covering ∼2 300 unique pedestrians. As discussed earlier, this benchmark set the standard for evaluating pedestrian detection. The established procedure for training is to use every 30th video frame which results in a total of 4 250 frames with ∼1 600 pedestrian annotations. Recently however, methods which benefit from more training data have resorted to a finer sampling of the videos (Chapter 5, Chapter 6, Nam *et al.* 2014, Zhang *et al.* 2015b), yielding 10× as much training data as the standard "1×" setting: ∼1 600 annotations on 42 782 frames. Here 10× and 1× refer to sampling every 3rd and 30th frame respectively. Detection methods are evaluated on a test set consisting of 4 024 frames. Typically, methods are evaluated under the so-called "Reasonable" setting, which excludes particularly hard to detect pedestrians from the evaluation. This subset consists of pedestrians that are taller than 50px and of which less than 35% is occluded. The provided evaluation toolbox additionally generates plots for different subsets of the test set based on annotation size, occlusion level and aspect ratio. We undertook a partial correction of the annotations for the analysis in Chapter 6, and these updated annotations have been adopted as a replacement in a lot of subsequent work.

The *KITTI Vision Benchmark Suite* (*KITTI*) (Geiger *et al.*, 2012) has a broader focus than the other datasets mentioned here. It covers several tasks relevant to autonomous driving beyond just detection, e.g. depth estimation, optical flow, scene flow, tracking, and semantic segmentation. The detection task covers both pedestrians and cars and includes 3D annotations. Similar to *Caltech*, the sequences were recorded in good weather in and around a single city — the mid-size city of Karlsruhe. It depicts both the inner city as well as surrounding rural areas and highways. However, Benenson *et al.* (2014) (supplementary material) show that it has different appearance statistics compared to the former. The training set contains 4 445 pedestrian annotations in 7 481 frames, and the test set consists of 7 518 frames with annotations withheld for evaluation.

CityPersons (Zhang *et al.*, 2017b) is based on the *Cityscapes* dataset we present in Chapter 7 and published in Cordts *et al.* (2016). The annotated part consists of 5000 high resolution images ($2048px \times 1024px$) sourced from videos recorded in 50 cities in and close to Germany. The original dataset provides pixel-level class and instance annotations for a variety of classes relevant for automotive applications. *CityPersons* focuses on pedestrian detection and extends the original annotations with bounding boxes that each cover the full extent of a pedestrian, requiring annotators to make a best guess when the pedestrian is occluded. Given that each annotated image is both part of a stereo pair and a short video clip, the dataset is suitable for methods that leverage both 3D and motion information. The additional pixel-wise labels enable methods that explicitly take semantic context into account for detection. Sequences were recorded in inner cities, often in crowded areas, resulting in denser images on average than other datasets in terms of the number of pedestrians.

While *Cityscapes* (and by extension *CityPersons*) were larger in scale and geographical spread than previous datasets, recordings were made during the day with less diversity in terms of weather conditions as well as location compared to subsequently released datasets. Several datasets have been compiled recently that cover more diverse and adverse conditions, with some upping the scale considerably.

The *KAIST Multispectral Pedestrian Dataset* (Hwang *et al.*, 2015) consists of daytime and nighttime sequences, with both RGB and thermal camera data. The *NightOwls* dataset (Neumann *et al.*, 2018) focuses on the night-time setting, providing 40 sequences comprising 279k frames covering different weather conditions and seasons. Sakaridis *et al.* (2018) present a method for simulating fog and apply it to *Cityscapes* resulting in *Foggy Cityscapes.*

Recent large-scale datasets include *EuroCity Persons* (Braun *et al.*, 2019), which was recorded during day and night in 12 European countries across four seasons. It includes 238K person instances annotated with both bounding boxes and orientation, making it relevant as well for pedestrian trajectory prediction. The *Berkeley DeepDrive 100K* dataset (*BDD100K*) (Yu *et al.*, 2020) consists of 100K video sequences from four different U.S. metropolitan areas, and provides annotations (one frame per sequence) supporting various semantic understanding tasks including object detection.

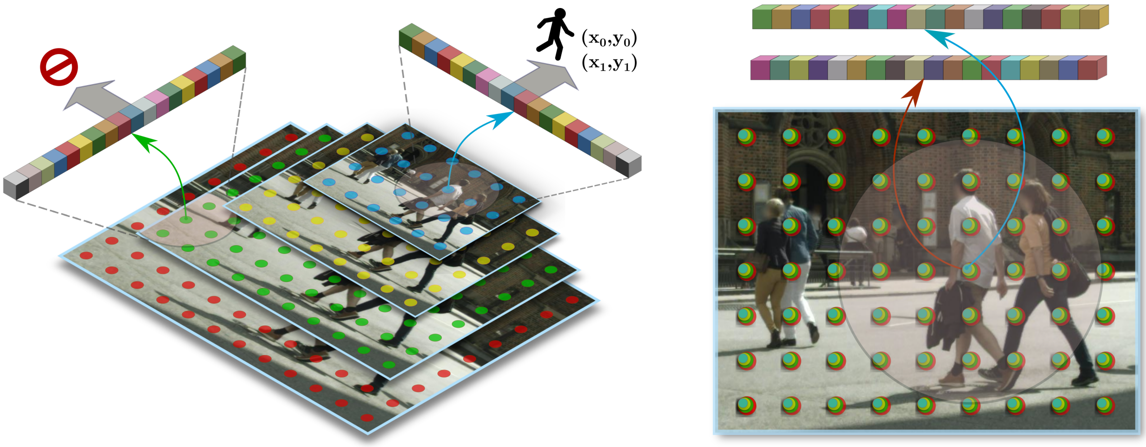
Most pedestrian detection datasets consist of recordings from a moving vehicle. This limits not only scene diversity but also scene density: Crowds tend to occur infrequently in the direct vicinity of the recording vehicle, unless it for example stops at a crosswalk. A couple of recent datasets thus target more general settings with the express goal of capturing more crowded scenes. *CrowdHuman* (Shao *et al.*, 2018) and *WiderPerson* (Zhang *et al.*, 2020b) both contain 10K-20K images with roughly 400K person instances each. This results in an average of 22.6 and 30 persons per image respectively, comparing quite favourably to other popular datasets, e.g. *Caltech* (0.32) and *CityPersons* (6.4).

Finally, Huang and Ramanan (2017) present the *Precarious Pedestrian Dataset*, which focuses on rare cases that are underrepresented elsewhere. They collect 951 images that depict potentially dangerous situations such as pedestrians texting or children playing in the street. Additionally, they use an adversarial generation framework to produce synthetic training data to complement the real images.

2.3 Methods

In this section, we cover recent work on pedestrian detection. In Chapter 4, we present a retrospective analysis of "classical detectors" from the period 2004-2014 — "classical" meaning detectors based on hand-crafted features. Here we focus mostly on recent advances in pedestrian detection since then, i.e. "modern" detectors that rely on deep neural networks (DNNs). Most of these advances have critically depended on more general research on representation learning and generic object detection. For an excellent comprehensive overview of the latter, we recommend the survey of Liu *et al.* (2020).

Since this thesis covers a very wide span of methods and since pedestrian detection cannot be discussed in isolation from generic object detection, we have opted to organise this section as follows. While a lot of progress has been made in detection judging by improvements in benchmark performance, fundamentally little has changed in the basic approach to this problem in the last couple of decades. At a high level, both



(a)

(b)

Figure 2.3: Object detection — especially the enduring sliding-window approach — can be viewed as labelling a multi-dimensional grid, where each grid point represents a 2D spatial location, an image scale, and potentially a shape prototype. The area surrounding each point is summarised with a feature vector and then either labelled as background or as an object with corresponding bounding box coordinates **(a)**. Many classical detectors involved explicit image rescaling to search for objects of different sizes, but most newer detectors extract features in a single pass and sample local representations as needed **(b)** — the overall approach, however, has fundamentally remained the same.

classical and modern detectors share a common scheme as well as similar design choices especially when it comes to efficiency. We will thus start by discussing these in Sec. 2.3.1. Naturally, there are design choices specific to modern detectors that follow from their reliance on end-to-end representation learning with deep neural networks. These we then outline in Sec. 2.3.2. Finally, we conclude the section with a detailed look at recent research in pedestrian detection. In Sec. 2.3.3, we focus on two points in particular: (i) the transition from classical pedestrian detectors to modern detectors, and (ii) current research trends in pedestrian detection.

2.3.1 Basic Approach

The very basic approach to object detection consists of the following elements: (i) deciding where to look in the image, (ii) at each selected location, extract a representation of the local content, (iii) decide if this represents an object and determine its precise extent, and finally (iv) aggregate all the local decisions into a coherent, non-redundant set.

This formula has survived the transition from classical to modern detectors, and most detectors follow one particular instantiation thereof: the "sliding window" paradigm. This means that the decision on where to look is made in advance and independently of the image content: namely at every point on a dense multi-dimensional evaluation grid that spans different image scales, different 2D locations on the image plane, and possibly different shapes. Each point on the grid represents a possible sub-image or image window, and detection can be viewed as assigning class labels to grid points, as well as bounding box dimensions if a point corresponds to an object (Fig. 2.3).

The naive implementation of the above, whereby each window is processed equally and independently, is very rare (Rowley *et al.*, 1995). It is neither feasible in most cases, nor is it even necessary: There are often different types of redundancies that can be exploited. A good way to introduce different detector variants is by discussing common design patterns that seek to overcome the intractibility of the naive sliding window approach. Most detectors use one or more of the following strategies:

- reducing the number of evaluations (sparse evaluation)
- spending more time on promising locations (cascade strategy)
- sharing the computational burden across multiple evaluations (feature sharing)

Since this grid is typically dense, each object will span multiple locations. Later, we will discuss how objects are commonly assigned to these grid points during training, as well as methods for handling the redundant detections that result at test-time due to the prediction density.

Sparse Evaluation

Some methods depart from the sliding window paradigm entirely, and explicitly or implicitly select a subset of locations to visit on the evaluation grid. Some leverage scene geometry, e.g. in the form of ground-plane constraints (Sudowe and Leibe, 2011) or via stereo information (Gavrila and Munder, 2007; Keller *et al.*, 2009; Benenson *et al.*, 2012). More commonly, promising locations are identified using low-level appearance cues, e.g. with interest point detection (Weber *et al.*, 2000), segmentation (Gu *et al.*, 2009), or more recently so-called object proposal methods (Hosang *et al.*, 2016). Some of these rely on hierarchically building up image segments that might contain objects (Uijlings *et al.*, 2013; Pont-Tuset *et al.*, 2017). Others sample windows based on saliency cues and score the “objectness” of these windows “objectness” (Alexe *et al.*, 2012). Previous top-performing detectors relied on this strategy, e.g. *Regionlets* (Wang *et al.*, 2013) and *R-CNN* (Girshick *et al.*, 2014). In the latter case, roughly 2000 boxes are extracted using *Selective Search* (Uijlings *et al.*, 2013) and classified with a CNN. A few methods rely on active search strategies: i.e. deciding where to look sequentially. The starting point is either the full image (Caicedo and Lazebnik, 2015; Lu *et al.*, 2016), a sparse set of proposals (Gonzalez-Garcia *et al.*, 2015; Mathe *et al.*, 2016) or a dense set of image windows from a low-resolution image and zooming in as needed (Uzkent *et al.*, 2020).

Cascade Strategy

While grouping-based or active search strategies for identifying a sparse set of object hypotheses are conceptually attractive, most detectors including the state-of-the-art rely on dense evaluation (Ren *et al.*, 2015). Denser evaluation grids have been shown to lead to better results for both classical (Dollár *et al.*, 2009a) and modern detectors (Lin *et al.*, 2017c). Many methods thus resort to a cascade strategy: Rather than carry out a full evaluation at every possible location, quickly rule out unpromising ones. This can take on many forms depending on the nature of the detector. Part-based detectors such as the *Deformable Part Model* detector (*DPM*) (Felzenszwalb *et al.*, 2010), which consist of a hierarchy of object and part classifiers, can exclude locations based on merely evaluating the coarse root classifier. Methods that rely on cascaded classifiers such as *AdaBoost* (Viola and Jones, 2004) can use early rejection thresholds to terminate evaluation. The most popular approach however is using detection-based object proposal methods. A dense, sliding window approach is used to identify promising subimages, which are then further processed. Most early CNN-based pedestrian detectors relied on classical detectors to provide hypotheses to a CNN-based refinement step. We present one such method in Chapter 5 and further analyse it in Chapter 6. State-of-the-art detectors on the other hand rely on integrated pipelines with an object proposal generator that shares features and is jointly trained with the subsequent refinement network, e.g. *Faster R-CNN* (Ren *et al.*, 2015).

Feature Sharing

Feature extraction is typically the most expensive part of detection, and correspondingly a lot of effort has gone into making this step efficient. The most effective hand-crafted feature representations consist of aggregations of local colour and edge features. Within a single image scale features can thus be reused for different windows. However, since these are shallow, local features, they're sensitive to image scale. Correspondingly, many classical detectors require dense image pyramids with up to 50 scales (Dollár *et al.*, 2009a; Benenson *et al.*, 2013). To mitigate the cost of image rescaling and feature recomputation, some methods instead resort to feature scaling: Extract suitable feature maps for a sparse set of scales and interpolate between them (Dollár *et al.*, 2014). An alternative approach that also uses a sparse image pyramid relies on multiple scale-specific models. These are applied together to each of the pyramid levels, effectively resulting in a dense scale evaluation. Benenson *et al.* (2012) combine the last two strategies to reduce the depth of the image pyramid even further: They use feature-scaling together with multiple scale-specific models.

While some early CNN-based detectors required the use of image pyramids (Sermanet *et al.*, 2013, 2014), recent detectors push feature sharing to its limits. They essentially do away with image pyramids entirely and instead rely on: (i) a deep hierarchy of increasingly complex feature maps extracted in a single pass over the image, i.e. a feature rather than an image pyramid, and (ii) several scale-specific classifiers that can be attached to different levels of the feature pyramid. These feature maps range from high-resolution low-level features (e.g. edges, colour differences) to low-resolution high-level features that correspond to semantic concepts such as objects, groups of objects and scenes even (Zhou *et al.*, 2015a). CNNs are essentially deeply nested filter banks that have the capacity to also specialise to the same object at different scales. The depth and capacity makes them less sensitive to scale than shallow hand-crafted features. A typical deep pedestrian detector can thus get away with considering fewer than 10 (virtual) image scales rather than 50. The distinction between the two approaches is illustrated in Fig. 2.3.

2.3.2 Modern Detectors

In this section, we will discuss some design choices that distinguish various CNN-based detectors. In the previous section, we described how such detectors make heavy use of feature sharing by extracting a feature pyramid — i.e. a collection of feature maps — from an image in a single pass. A key distinguishing factor is thus the feature extraction network (often referred to as the backbone). This is typically a network designed for single-object recognition (Krizhevsky *et al.*, 2012; Simonyan and Zisserman, 2015; He *et al.*, 2016) which prior to classification produces low-resolution feature maps at the highest level through the repeated use of sub-sampling operations.

Since detection also requires fine-grained spatial information, another key element of modern detectors is addressing this loss of resolution. Popular solutions include the use of dilated convolutions that avoid some loss of detail (Chen *et al.*, 2015a; Yu and Koltun, 2016), introducing top-down connections that merge high-level and low-level feature maps and recover lost detail (Lin *et al.*, 2017b; Tan *et al.*, 2020), or designing the network from the ground up to maintain high-resolution representations throughout (Pohlen *et al.*, 2017; Wang *et al.*, 2020).

Once we have a feature pyramid, how do we detect objects? In Sec. 2.3.1, we presented detection as the problem of labelling a multi-dimensional grid, which represents the search for objects across different spatial locations and scales. An important question is then how to map this grid to the feature pyramid, or in other words how to sample feature vectors for grid points that correspond to different locations and scales. Since the feature maps are registered to the input image, establishing spatial correspondence is relatively straightforward. Scale handling, on the other hand, is challenging, especially since we typically have a small number of feature maps that need to cover a large range of object scales. Another distinguishing factor between detectors is then the assignment of different scale ranges to different feature maps. Some assign the full range to the final feature map (Ren *et al.*, 2015), and others spread these out across the feature pyramid (Liu *et al.*, 2016).

This is where the other main component of most detectors — besides the feature extraction network — comes in: a set of classifier-regressor pairs responsible for identifying and precisely localising objects. Each pair is responsible for a disjoint set of grid points, meaning that they are trained to specialise to a separate range of object scales and sometimes shapes, and sometimes even spatial locations (Redmon *et al.*, 2016). Classifier-regressor pairs often specialise to narrower scale ranges than the ranges assigned to one level of the feature pyramid, and thus sometimes several are attached to the same feature map (Ren *et al.*, 2015).

How do classifier-regressor pairs specialise to different scales? Alternatively, how do we assign objects to points on the evaluation grid? There are broadly speaking two approaches to this: anchor-based and anchor-free methods. In the anchor-based approach (e.g. Ren *et al.* 2015, Zhang *et al.* 2017b), each grid point is associated with a so-called “anchor”: a template bounding-box that represents the default object size and shape for that point. Any object whose bounding box sufficiently overlaps with an anchor (measured via intersection-over-union) is assigned to that point and correspondingly to the responsible classifier-regressor pair. Objects can be assigned to multiple points on that basis. The regression target is then often the offsets between the anchor coordinates and the target bounding box.

In contrast, anchor-free approaches (e.g. Tian *et al.* 2019) assign an object to a grid point if the latter is within a certain distance to the object centre. These also often require different regression targets compared to anchor-based approaches, e.g. distance to the sides of the bounding box (Tian *et al.*, 2019) or the scale of the bounding box (Liu *et al.*, 2019d). (Zhang *et al.*, 2020a) compare the two approaches on equal footing and demonstrate that this assignment process matters more for performance than the regression target.

Since each object can be assigned to multiple grid points, detectors are trained to make redundant detections. Post-processing is thus necessary to reduce these to a coherent set. The most common approach is referred to as greedy non-maximum suppression and involves a simple strategy of eliminating lower-confidence detections that cover an area already explained by a higher-confidence prediction. We discuss this approach together with some alternatives towards the end of the chapter.

The final design choice we will discuss here is the number of detection stages. So far, we assumed only one in our discussion, but many detectors use two stages and are based on the popular *Faster R-CNN* detector (Ren *et al.*, 2015). The first stage produces class-agnostic object hypotheses: The backbone network extracts a feature pyramid, and multiple classifier-regressor pairs specialise to different object shapes and sizes but only distinguish between a generic object class and background. In the second stage, fixed-length feature vectors are sampled for each surviving object hypothesis and are processed by further classification and regression subnetworks. Two-stage detectors typically outperform single-stage approaches in part because the feature re-sampling of the second stage essentially normalises objects to a common scale, making it easier to model their appearance.

2.3.3 Recent Research

So far, we discussed general aspects of object detection. Now, we summarise recent research on pedestrian detection. We start by describing early CNN-based approaches which mostly focused on classification, relying on separately trained detectors to provide object hypotheses. Several of these resort to part-based deep mixture models similar to *DPM*, or on context modelling. In this setup, complicated domain-specific models do not show an advantage over generic classification networks as we show in Chapter 5.

Next, we discuss the adaptation of end-to-end detectors to pedestrian detection. The key to this adaptation is careful scale handling at the input level, feature level and classifier level. Additionally, we describe examples of end-to-end ensembles and cascaded detectors.

Another line of research focuses on reformulating the localisation objective, e.g. replacing bounding box coordinate regression with keypoint localisation or other equivalent targets. Similarly, some methods show that useful weak pixel-level supervision can be derived from bounding boxes or other data modalities.

Several end-to-end detectors also consider explicit part and/or occlusion handling. Common approaches include occlusion-aware loss functions, part-aware feature pooling and reweighting, as well as specialised classification/regression branches. Finally, we cover a few works that attempt to address the problem of suppressing redundant detections.

Mixed Pipelines

Early NN-based pedestrian detection methods with few exceptions (e.g. Sermanet *et al.* (2013)) relied on a mixed cascade strategy: Hypotheses supplied by a simpler, classical detector are rescored by a separate neural network. This is in contrast to modern detectors which process a full image end-to-end, sharing computations across hypotheses. Different variants of mixed approaches include: (i) operating directly on the detector scores (Ouyang and Wang, 2012; Ouyang *et al.*, 2013), (ii) merely learning a classifier on top of hand-crafted features (Zeng *et al.*, 2013), (iii) learning both a feature extractor and a classifier that process hand-crafted features (Ouyang and Wang, 2013a; Luo *et al.*, 2014), all the way to (iv) end-to-end classification networks that operate on RGB inputs (Tian *et al.* (2015a); Li *et al.* (2018), and the method we present in Chapter 5).

Some of these methods incorporate problem-specific modelling, mainly inspired by part-based approaches such as the *DPM* detector mentioned above (Felzenszwalb *et al.*, 2010). (Ouyang and Wang, 2013a) include a layer designed to capture part deformation costs, and (Luo *et al.*, 2014) explicitly model pedestrians with mixtures of full body and part templates — or “components” in *DPM* parlance. These aim to cover the appearance variation of certain object classes that can result from commonly occuring viewpoints and poses (e.g. “frontal view” vs. “side view” or “standing” vs. “sitting”), and are at least as important as capturing part deformations (Divvala *et al.*, 2012). Tian *et al.* (2015a) demonstrate a conceptually simpler — if more computationally demanding — approach to modelling pedestrians as collections of parts. They learn part-specific CNN detectors that each focus on one rectangular area of the bounding box out of a pool of 45. An SVM is used to whittle these down to the most relevant ones.

However, vanilla CNNs have the structure and capacity to capture these intra-class variations and parts automatically (Zhang *et al.*, 2018c). Different filters can specialise to different "components", and max-pooling layers for example allow for some robustness to deformation. Our results in Chapter 5 also provide some evidence for this. We find that a simple classification network already outperforms a specialised model (Ouyang and Wang, 2013a) when provided with the same training data and object hypotheses at test time. Along the same lines, Li *et al.* (2018) demonstrate strong results with an ensemble of two scale-specific, but otherwise generic classification sub-networks on top of the feature extraction network. The classifier scores are fused with weights determined by the height of the proposal.

Integrated Pipelines

Most methods discussed up until this point rely on object proposals from a classical pedestrian detector operating on hand-crafted features. This verification approach is slow and dependent on the quality of the proposal stage, which is also trained separately.

After end-to-end NN architectures became the norm for generic object detection, e.g. *Faster R-CNN* (Ren *et al.*, 2015), such integrated approaches only started to outperform mixed pipelines for pedestrian detection with some delay. What was holding integrated pipelines back? In a nutshell, the bottleneck was inadequate scale handling. Detectors such as *Faster R-CNN* were designed against benchmarks like *ImageNet* or *PASCAL VOC*, where the distribution of object sizes is a much narrower one compared to pedestrian datasets (see Chapter 7). What it took to adapt such detectors was simply: (i) higher-resolution inputs, (ii) higher-resolution features, and (iii) the appropriate set of scale-specific classifiers (Zhang *et al.*, 2016a; Cai *et al.*, 2016; Zhang *et al.*, 2017b).

Zhang *et al.* (2016a) inverted the usual pipeline of classical detector followed by CNN-based classification. Their method *RPN+BF* involved a boosted decision forest verifying proposals from a modified *Region Proposal Network* (*RPN*) (Ren *et al.*, 2015). *RPN* was adapted in several ways: higher input resolution, anchors covering a larger scale range with a default aspect ratio of 0.41, and dilated convolutions Chen *et al.* (2015a); Yu and Koltun (2016) for higher resolution feature maps. The second stage of *Faster R-CNN* was discarded as it was found to hurt results.

Concurrently, Cai *et al.* (2016) addressed the same problems but by carefully adapting *Faster R-CNN* rather than resorting to a mixed pipeline. Similarly, they increase the input resolution as well as the final feature map resolution (but with upsampling rather than dilated convolutions). More critically, the specialised classifiers for anchors of different sizes are assigned to different layers of the network. This is a common strategy for single-stage detectors such as *SSD* (Liu *et al.*, 2016), but not so much for two-stage detectors. With these improvements to *RPN*, they find that the second stage improves performance significantly in contrast to the findings from Zhang *et al.* (2016a).

(Zhang *et al.*, 2017b), besides presenting a new benchmark based on the dataset we describe in Chapter 7, also successfully modified *Faster R-CNN* to obtain strong pedestrian detection performance. Similar to the aforementioned works, they upscale the input image and increase feature map resolution by removing a sub-sampling layer. Additionally, they propose to use several scale-specific anchors but based on the training set statistics.

All three of the above methods perform similarly well on the *Caltech* "Reasonable" test set, but *MS-CNN* outperforms the others by a significant margin on the occluded subsets, but it's hard to say which element is responsible. Incidentally both *RPN+BF* and *MS-CNN* rely on explicit hard negative mining, which was critical for classical detectors but is less commonly used with modern ones.

While proper scale-handling is one important aspect of detection, another powerful and recurring element is the cascade as mentioned above. Two-stage detectors such as *Faster R-CNN* are a specific form of cascade, whereby in the second stage a fixed-length representation is sampled for each hypothesis and processed separately. Other works have explored variations on the cascade idea with single-stage pedestrian detectors Liu *et al.* (2018); Brazil and Liu (2019).

With *Faster R-CNN*, positive training examples are assigned to locations in the first stage with a more permissive IoU criterion than in the second. Liu *et al.* (2018) argue that this successive refinement is more important than the feature resampling between stages. They thus propose to stack multiple predictors on top of each other, trained with a successively stricter assignment of hypotheses. Multiple such cascades are attached to different levels of the feature pyramid. Brazil and Liu (2019) also propose to use multi-level cascades, but additionally treating them as a form of ensemble. Intermediate predictions from one cascade are fed to the others.

Alternate Objectives

The above methods involve detectors trained to classify a bounding box and regress to its adjusted coordinates. Several works recently either: (i) reformulate the task to localise keypoints on the pixel grid rather than predict continuous bounding box coordinates, or (ii) include complementary targets.

Bounding box annotations are typically produced by marking the top-left and bottom-right corners of the box. In Chapter 6, we argue that marking the top of the head and the midpoint between both feet is easier, and results in more consistently aligned bounding boxes which benefits performance. Song *et al.* (2018) propose to train a network to label these two points as well as the line connecting them. A Markov Random Field is used to group pairs of keypoints into detections. Instead of two points per detection, Liu *et al.*

(2019d) propose to predict the centre point and pedestrian scale, which are sufficient to generate the corresponding bounding box without an additional grouping step.

These works can be viewed as part of a trend in generic object detection. Law and Deng (2018) for example detect the upper-left and lower-right corners of the bounding box and Zhou *et al.* (2019a) detect the centre and extreme points. Extreme points are points that lie on the bounding box as well as the object boundary, and they predict one such point per bounding box side. While this requires annotated segments, it results in an easier detection task when an object is irregularly shaped and the bounding box corners are relatively distant from its boundary.

Rather than replace the standard detection objective, other approaches augment it. Mao *et al.* (2017) show that performance can be improved by either providing the network with additional feature channels beyond the RGB image (e.g. edges or optical flow) or by training it to predict these quantities. Similarly, Xu *et al.* (2017) train a detector to additionally predict the thermal image corresponding to the RGB input. This, however, requires additional recordings to obtain the extra target.

Lin *et al.* (2018), Brazil *et al.* (2017) and Noh *et al.* (2018) all demonstrate that weak pixel-level supervision derived from bounding boxes is beneficial. If additional pixel-wise semantic annotations are available, these can also help as previously shown by Tian *et al.* (2015b) and Costea and Nedevschi (2016).

Luo *et al.* (2020) use generative model trained on synthetic data to hallucinate a birds-eye view map from a frontal street image. A second module localises people in the hallucinated map which indicates their scale in the frontal image. This informs the detector as it is applied to the frontal image.

Part Modelling and Occlusion

One aspect of pedestrian detection that an increasing number of methods focus on is occlusion. Some methods focus on improving the detection loss functions. Wang *et al.* (2018c) posit that intra-class occlusion is a bigger problem than inter-class occlusion. They then propose an expanded bounding box regression loss, which besides encouraging predictions to match their targets, has two “repulsion” terms: one that penalises overlap between predictions assigned to different targets, and another that penalises overlap between predictions and unrelated targets. Similarly, Zhang *et al.* (2018b) propose an “aggregation loss” that forces predictions assigned to the same target to cluster.

Other methods present approaches that involve some form of part-modelling, albeit one that doesn’t require additional part annotations: merely a bounding-box annotation for the visible part of the pedestrian. Zhang *et al.* (2018b), besides the aggregation loss, also address occlusion by pooling features not only from the full template, but also from five different parts. Visibility for each part is estimated and the resulting visibility scores are used to compute a weighted combination of these pooled feature vectors for the final decision. Noh *et al.* (2018) and Wang *et al.* (2018a) also predict part visibility. They divide the pedestrian template into rectangular grid cells, each representing a “part”. The latter also uses LSTMs to communicate information bidirectionally between them, with the motivation being that visible parts should boost scores of less visible ones. Zhang *et al.* (2018c) observe that different feature channels are sensitive to different parts, and learn to predict a vector that reweights feature channels. This amounts to a flexible part model that attends to visible parts in a dynamic manner.

Several works propose to use use specialised detection branches for different parts of the body. Zhou and Yuan (2018) propose a two-branch *Faster R-CNN*, with one branch for full-body prediction and another that regresses to the visible bounding box, both trained to be complementary. Huang *et al.* (2020) propose to do the same, but additionally use these separate predictions for non-maximum suppression, since the visible bounding-box is more appropriate for crowded scenes. Zhu *et al.* (2020) use separate predictions for head and full body for the same purpose, but this requires additional bounding-box annotations for the head.

Non-Maximum Suppression

Above, we listed several methods that propose additional detection targets or objectives such that the detector output becomes more useful for non-maximum suppression. These methods have largely stuck to the well-worn *Greedy NMS* procedure: Detections are sorted by confidence and selected in that order. A detection is suppressed if its IoU with any previously selected detection exceeds some fixed threshold. In the following, we discuss a few methods that aim to replace *Greedy NMS* itself.

First, two methods not specific to pedestrian detection: *Soft-NMS* (Bodla *et al.*, 2017) and *GossipNet* (Hosang *et al.*, 2017). *Soft-NMS* is a simple extension to Greedy NMS. When some detection A is selected, the scores of all detections that overlap with A are decayed based on the degree of their overlap. Hosang *et al.* (2017) propose to use a network, *GossipNet*, that operates on all detection hypotheses in an image. The network is trained to update the scores of these detections through pairwise comparisons in multiple stages such that one detection per object remains. Similarly, Liu *et al.* (2019c), add a sub-network to the detector which predicts an adaptive NMS threshold for each detection. The motivation is that a strict threshold is needed in sparsely populated areas of the image, but a more permissive one in more crowded parts.

Lee *et al.* (2016) observe that *Greedy NMS*, when choosing between two competing detections A and B, only makes the decision based on comparing detection scores, e.g. A looks more like a pedestrian than B. Instead, they argue that NMS should consider the following criterion: if B also looks like a pedestrian, does it look sufficiently different from A? To this end, they propose to use Determinantal Point Processes (DPP). For all N detection candidates, an $N \times N$ similarity matrix is set up. A DPP helps select the subset of these detections that results in the matrix with the maximum determinant, i.e. such that unary terms (diagonal) are high and similarity terms (off-diagonals) are low.

2.4 Summary

In this chapter we reviewed relevant material for pedestrian detection. We presented the task definition as well as relevant benchmarks and evaluation metrics. In the methods section, we started by describing basic patterns underlying both classical and modern detectors: (i) feature sharing, (ii) classifier cascades, and (iii) dense vs. sparse evaluation. Modern CNN-based detectors typically rely on efficient feature sharing as well as dense evaluation. These also differ in many respects, which we summarised next: (i) if and how low-level to high-level feature maps are combined into a feature pyramid, (ii) how locations in the feature pyramid are marked as positive or negative targets during training, (iii) how many classifier-regressor pairs are used and which parts of the feature pyramid and object space they operate on, (iv) whether a one- or two-stage approach is used, and (v) the localisation objective beyond mere bounding box coordinate regression.

Finally, we concluded our review with an up-to-date survey of recent research into pedestrian detection. Here, we focused on the following points: (i) early efforts to use CNNs for pedestrian detection, (ii) the transition from mixed pipelines — which used CNNs for verifying sparse object hypotheses supplied by classical detectors — to fully end-to-end detectors thanks to careful scale handling at multiple levels, (iii) specialised ensembles and cascaded detectors, (iv) the use of weak pixel-level supervision, (v) keypoint-based approaches, (vi) part-based modelling and occlusion handling, as well as (vii) non-maximum suppression.

In the final chapter, we will discuss this review in light of the results we present in chapters Chapters 4 to 6 and present possible future directions.

3 Related Work: 3D Human Shape and Pose Recovery

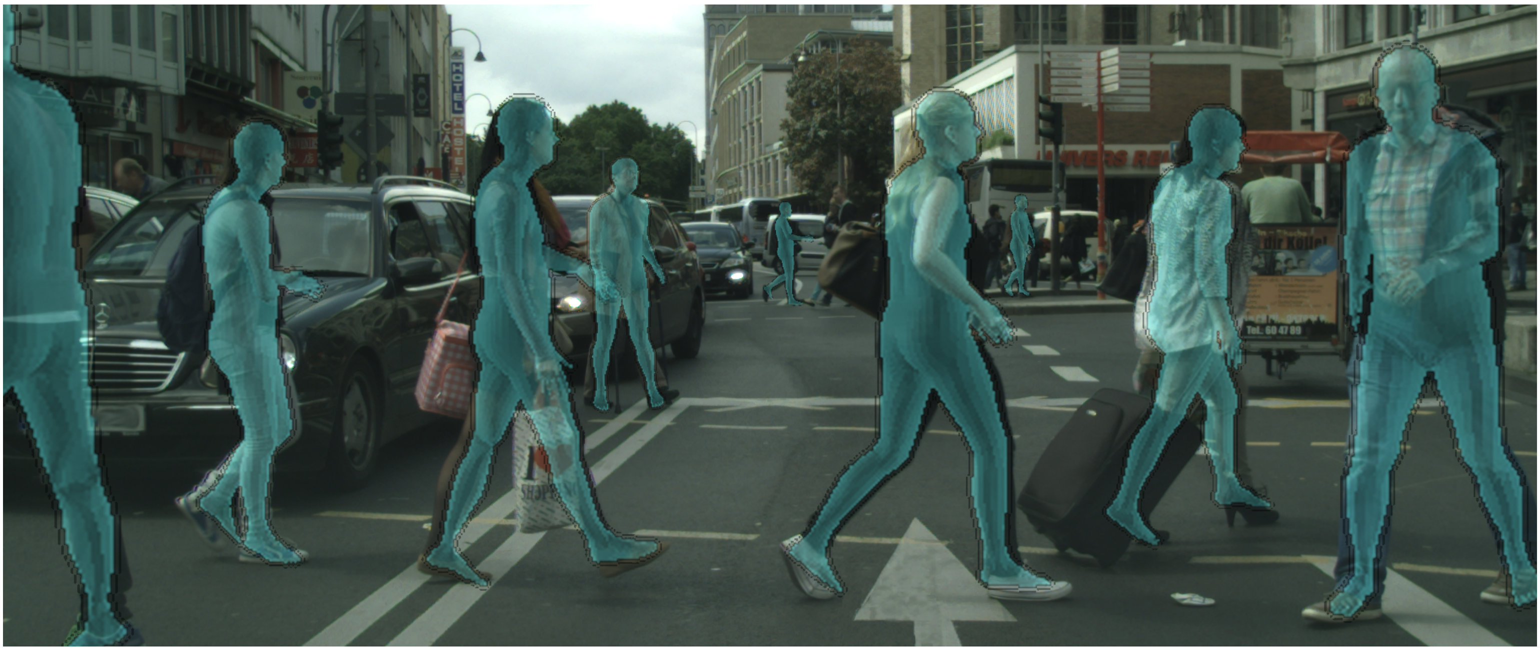


Figure 3.1: 3D human shape and pose recovery is a challenging task that requires us to recover a representation of body pose and surface, as this can help us e.g. determine where pedestrians are likely to move next. (image from *Cityscapes* (Chapter 7), results generated with method described in Chapter 9)

In the previous chapter we focused on localising people in images. Now we turn to the task of extracting richer human representations from single monocular images, namely 3D pose and shape.

Pose is typically taken to mean the locations of a set of body parts each represented as a point in 2D or 3D space, or alternatively the relative orientations of body parts in space, with each part represented as a virtual "bone". Pose, despite being a sparse representation of the human body, encodes a lot of information relevant to interacting with and understanding humans. It can encode simple gestures, activities and even certain emotional states.

We are also interested in extracting shape, i.e. some representation of the body surface. While pose is a practical and informative representation, it often abstracts away valuable information, including subtle social signals and precise interactions with objects and the environment. There are also perceptual considerations that make surface representations desirable: Skeletons are almost always indirectly inferred, whereas the surface of the body is often — at least partially — observed in the image. This makes it possible to verify the surface estimates against other observable quantities such as image edges and depth.

Pose and shape are naturally intertwined not least because our body surface deforms as we change our pose, so it makes sense to address these tasks together. Previous work has demonstrated that even when the goal is to merely to extract the body surface of humans from 3D point clouds — in principle easier than from monocular images — one can benefit from first explicitly extracting pose to constrain the surface extraction.

Challenges

Most of the challenges associated with localising people in images also apply to shape and pose recovery. There are a host of others that are more specific to this task. These include ambiguities resulting from projection and missing information due to self-occlusion and clutter. The output space is also difficult to define and capture flexibly and efficiently. Ground truth acquisition in natural settings is also very challenging.

The challenges start with defining the output space itself, which is anything but straightforward. Skeletons are a natural choice of representation for pose, but the choice is less obvious in the case of the body surface. There is a trade-off between fidelity to surface detail and efficiency of representation.

But even just the space of poses, which can be captured by a relatively low-dimensional skeleton, is rather complicated to navigate. Many possible values in this space will correspond to implausible or even impossible poses. This affects methods that learn to predict pose in a discriminative, data-driven manner, as well as model-based approaches that search for a configuration that explains evidence from the image.

Given that the human body is highly articulated, often one also has to contend with missing data due to self-occlusion. Occlusions from other people and the environment result in additional ambiguity. Sometimes resolving these ambiguities will necessitate jointly reasoning about fine-grained appearance cues, such as shadows and lighting (Balan *et al.*, 2007a) or the type of activity being carried out (Luvizon *et al.*, 2018), but also about the geometry of the environment (Hassan *et al.*, 2019). When we try to perceive 3D pose and shape from 2D images, we have to deal with the projective ambiguity, namely the fact that many different 3D poses can result in a similar 2D projection.

Ground truth acquisition is also challenging. Acquiring data in natural settings without nuisance signals such as visible markers is already challenging enough. Adequately covering the space of possible poses makes this even more challenging. This also means that domain adaptation and weakly supervised learning play a larger role here. It's much easier to record 3D data in a studio or to generate synthetic data, but difficult to transfer what is learned from such data to everyday scenes. It's also much easier to obtain 2D data, so many methods resort to learning about humans in 3D without explicit 3D supervision.

Many methods assume that there is a single person that has been pre-localised, e.g. with a bounding box-based detector. All the above difficulties apply in this restricted setting, but the problem becomes even more difficult when dealing with multiple people who may even be interacting with each other.

Summary

In this chapter, we will mostly focus on 3D pose and shape estimation research in the last five years, with the occasional nod to older work as needed. We will mostly restrict our discussion to methods that operate on single, monocular images, as this is the task we address in the last part of this thesis. While we are ultimately interested in recovering shape and pose, much of the relevant work — especially on benchmarking — focuses solely only on 3D pose, i.e. recovering a skeleton rather than a surface representation. Our discussion will accordingly cover this work as well. We will refer to methods that recover some three-dimensional structural description of the human body as *3D human body recovery* (*3DHBR*) methods. This subsumes all methods we cover here.

We will follow a structure similar to that of the previous chapter. We will first define the task as well as relevant evaluation metrics. We will then describe notable datasets with a focus on advances in dataset acquisition. Unlike for pedestrian detection, obtaining consistent ground truth annotations for arbitrary images is itself a challenging research question and we will describe efforts to address it. We then turn our attention to recent methods. Methods in this area are diverse and difficult to categorise neatly, so we will organise a large part of the discussion around two questions: (i) What parametrisations of pose and shape are conducive to learning and inference? (ii) What constraints can be applied to encourage valid outputs either during training or inference? We will then separately discuss two classes of methods: 3DHBR for multiple people, and neural network-based methods that incorporate rich statistical body models. The last part of this thesis (Chapter 9) presents a method in the latter group.

3.1 Task Definition

The task of extracting 3D shape and pose is not a well-defined one, as there are many ways to parametrise these quantities also at different levels of granularity.

The most common model of human pose is the skeleton, a tree-structured collection of keypoints mostly corresponding to articulated joints (e.g. knees, elbows) or surface locations (e.g. eyes, nose). Attached to each keypoint is its location, either in pixel or in 3D space. In the case of 3D, this is typically relative to some pre-defined root keypoint. Skeletons encode some information on shape, e.g. limb lengths, and significant — if incomplete — information on body pose. Depending on the collection of keypoints under consideration, some pose information will not be encoded: such as axial limb rotations and head rotation about the longitudinal axis. With the exception of very recent work (Pavlakos *et al.*, 2019a; Martinez *et al.*, 2019; Weinzaepfel *et al.*, 2020; Choutas *et al.*, 2020), most pose estimation methods — whether 2D or 3D — ignore the fine-grained articulation of hands and feet, typically stopping at the wrists and ankles.

In this work, we are interested in extracting a richer output representation, both in terms of pose and shape. A richer pose representation would encode not just joint locations but full limb rotations as well, and a richer shape representation would capture the surface of the body, not just limb lengths. This naturally begets the question: How do we model the surface of the human body?

3.1.1 Parametric Models of 3D Pose and Shape

There are many representations of the human body shape in the literature. While a full discussion is beyond our scope, we will cover some relevant aspects to our work.

Popular surface representations include point clouds, voxels, implicit surfaces, meshes as well as hybrid representations. Here we need to distinguish on the one hand between the surface representation itself, and on the other hand whether the range of allowable configurations is further constrained by some underlying parametric model. The relative advantages and disadvantages of the aforementioned surface representations also depend on the use case. Here, this is inferring 3D structure from 2D data with a view towards understanding: where a person might be going, what they might be doing, etc.. Accurate reconstruction or generation of synthetic human models may impose different requirements.

Meshes for example are very flexible and powerful representations, but their high fidelity comes at the cost of high dimensionality and unwieldiness when it comes to handling topological changes. Predicting vertex locations and their connectivity is highly non-trivial without any limiting assumptions. This poses challenges in the case of general objects and surfaces, but luckily in the case of humans — a limited class of shapes — there are statistical regularities we can exploit, making meshes suitable as an underlying representation for understanding images of people.

For our purposes, we thus opt for a class of articulated body models, which parametrise a high-dimensional body surface mesh in terms of separate, lower-dimensional shape and pose representations, e.g. the *SMPL* model (Loper *et al.*, 2015). Pose is represented as the rotations of a set of connected body parts that make up a skeleton. Joint locations are a function of body shape, and the surface of the body deforms rigidly and non-rigidly as a function of the joint angles. The space of shapes is spanned by a small number of basis shapes, and so the shape representation consists accordingly of a small number of basis weights.

The benefits of such parametric models are legion: They decouple identity-dependent shape from articulated pose. Location is automatically decoupled from the representation as well, unlike with e.g. voxel representations. The surface is efficiently modelled by exploiting the statistics of human body shape as well as the regularities of how it deforms as one moves. As such, they capture a lot of prior knowledge about the human body and can be fit to sparse data (Bogo *et al.*, 2016). They are also generic enough to describe the shape and pose of a wide variety of people in a wide variety of poses. Semantic correspondences are also built into such models, and there are a number of ways to match them against observations in the image, e.g. body keypoints and silhouettes, via efficient model abstraction.

Some such models, specifically the one we make use of in this work, make a number of mild simplifying assumptions: There is such a thing as rest shape (sometimes referred to as identity-dependent shape), i.e. the shape of a person in some canonical static pose. Of course "rest shape" is not fully identity-dependent as it can depend on a variety of things, e.g. rate of breathing, how much one has just eaten, injuries, or prior activities. Another assumption is that rest shape will deform both rigidly and non-rigidly purely as a function of instantaneous pose, but of course in reality speed of motion matters as well — w.r.t. body fat for example. However, these and similar assumptions are not particularly limiting for our purposes and such models remain powerful and expressive.

Early parametric 3D models were based on simple geometric primitives, e.g. Metaxas and Terzopoulos (1993); Gavrila and Davis (1996); Sidenbladh and Black (2001); Plänkers and Fua (2001); Sigal *et al.* (2004). Eventually, statistical mesh-based models were learned from large databases of scans. These were richer than their predecessors but retained the low-dimensional representation. A key problem here is how to deform the surface as a function of shape and pose. Initial attempts focused on polygon deformations (Anguelov *et al.*, 2005; Hasler *et al.*, 2009). This was motivated in part by the transferability of such deformations across body sizes. On the other hand, this required costly optimisation to realign the triangles and recover a watertight mesh after applying shape and pose deformations.

Later models, most notably *SMPL* (Loper *et al.*, 2015), used linear blend skinning, which applies deformations via vertex displacements as a function of joint displacements. Pose-dependent corrective shapes are added to compensate for the artifacts that result from naive linear blend skinning. As the operations involved are linear functions of a small number of parameters, these models easily lend themselves to optimisation or embedding in neural networks as we will demonstrate later. Recent variants of this class of models have made them more expressive for faces and hands (Joo *et al.*, 2018; Pavlakos *et al.*, 2019a), suggested improvements in terms of model learning (Xu *et al.*, 2020) and parameter efficiency (Osman *et al.*, 2020).

3.1.2 Evaluation Metrics

Prior to the availability of standardised benchmarks, methods would be evaluated quantitatively in a variety of ways (Sigal *et al.*, 2010). With the introduction of the *HumanEVA-I benchmark*, the common evaluation metric became "mean per-joint error" or MPJPE for short. This is simply the average euclidean distance in terms of millimetres between predicted and ground truth joints, but with both sets aligned at a common root. One popular variant on this is the so-called "reconstruction error", in which prediction and ground truth additionally undergo a rigid body alignment which removes discrepancies in global rotation and scale, but sometimes only in scale.

Another metric which is less often used but has been argued for by Ionescu *et al.* (2014) and Mehta *et al.* (2017a) is measuring the percentage of correct 3D keypoints (3D-PCK). A keypoint is considered successfully detected if it is within 150mm of the ground truth joint, i.e. roughly half the size of a human head. This is analogous to the PCKh metric in 2D pose estimation (Andriluka *et al.*, 2014). Additionally, one can measure 3D-PCK for a range of thresholds and calculate the area under the curve (AUC). Unlike MPJPE, this metric is robust to small imperfections in the annotations.

Some methods that output a full mesh additionally report per-vertex error and segmentation accuracy. Per-vertex error is informative, as it can happen that a method correctly outputs keypoint locations but not limb rotations. Many approaches resort to keypoint projection losses to learn from 2D annotations, and this can lead to such errors that visibly affect the mesh but not the skeleton. Segmentation accuracy is similarly useful. The *LSP* and *LSP-extended* datasets (Johnson and Everingham, 2010, 2011) were annotated with six body part labels by Lassner *et al.* (2017), and mesh recovery methods report pixel-wise accuracy as well as $F1$-score.

For multi-person pose estimation, no special metrics exist. 3D-PCK is simply computed for all subjects individually and averaged per sequence (Mehta *et al.*, 2018).

3.2 Datasets and Benchmarks

For a long time the evaluation of 3D human body recovery (3DHBR) was a mostly qualitative affair. Methods were applied to a handful of images or sequences and the results analysed visually (Sigal *et al.*, 2010). Qualitative evaluation was typically restricted to either synthetic datasets (Agarwal and Triggs, 2004) or sequences that were not publicly available (Balan *et al.*, 2007b). While this was the case for other computer vision tasks as well, it took some time for this area to catch up.

With tasks such as object detection or 2D pose estimation — especially when involving monocular images — the act of annotating itself tends to be the easy part. While obtaining consistent annotations across images is certainly difficult, challenges are more likely to arise when the annotation effort needs to be scaled up, which can be addressed via crowd-sourcing with quality control (Johnson and Everingham, 2011; Su *et al.*, 2012) and human-machine collaboration (Russakovsky *et al.*, 2015b; Benenson *et al.*, 2019).

With 3D shape and pose on the other hand, already obtaining a single ground truth annotation presents obstacles. This is especially the case if one requires ground truth recorded both in natural settings and in a manner that does not introduce nuisance signals to the image, e.g. visible markers on the body surface. Fortunately, a lot of work has gone into overcoming these difficulties and there is no shortage of challenging and unsolved datasets.

Since we are interested in recovering a 3D description of the body from monocular images and measuring our ability to do so, the discussion in this subsection will mostly focus on datasets of RGB images paired with 3D ground truth. However, stand-alone 3D data, e.g. body surface measurements recorded from range scanners and skeletal data obtained from motion capture systems, also play a pivotal role for many techniques.

Anthropometric data captured using range scanners has been critical for realistically modelling the space of human body shape. It has also enabled the development of statistical body models that jointly model shape and pose, which we described in Sec. 3.1.1. An early dataset that played a critical role in this area is *CAESAR* (Robinette and Daanen, 1999).

As mentioned above, synthetic datasets often make use of motion capture data to represent human movements realistically. The most widely-used source of such data is the (CMU Graphics Lab Motion Capture Database). Some methods rely on such data to learn priors on human pose and/or motion. Motivated by the limitations of then existing datasets in terms of the activities and range of motion they cover, Akhter and Black (2015) put together the *PosePrior* dataset. This dataset captures trained athletes carrying out various stretching poses. Recent developments have enabled the extension of diverse databases of pose and motion to also to cover articulated shape in a unified manner, e.g. the large-scale *AMASS* database (Mahmood *et al.*, 2019). This has been made possible by the development of statistical body models such as *SMPL* , methods to fit the former to a sparse set of markers such as *MoSH* (Loper *et al.*, 2014) and extensions.

First standardised benchmarks

The first standardised benchmarks for the task of 3DHBR were *HumanEVA-I* (Sigal *et al.*, 2010) followed by *Human3.6M* (*H36M*) (Ionescu *et al.*, 2014). Both consist of multi-camera sequences of actors performing everyday actions in a studio, and both use a combination of motion capture (mocap) and software to synchronise video and motion data. *HumanEVA-I* established a common evaluation procedure (more on that below) and *H36M* scaled things up terms of number of actions as well as the amount of data. The latter thus functions as a source of training data, and remains the de facto standard benchmark for 3DHBR, but this is starting to change due to: (i) the increasing interest in tackling 3DHBR “in the wild”, as well as (ii) advances in both data acquisition and 3DHBR itself that make the former possible.

Data acquisition in natural settings

Recently, a number of datasets have been released that use a mixture of non-invasive sensors and creative techniques to obtain ground truth data in outdoor settings: For their *MPI-INF-3DHP* dataset, Mehta *et al.* (2017a) apply a commercial marker-less mocap system to sequences recorded both outdoors as well as indoors. A subset of the sequences are recorded against a green screen, allowing for the compositing of actor footage against natural backgrounds with some appearance augmentation. A total of 1.3M frames from 14 cameras are provided, including 500k images from 5 chest height cameras. They also show that this yields accurate 2D annotations such as keypoints.

While the use of marker-less motion capture allows for the recording of sequences in more natural settings, the system employed for the *MPI-INF-3DHP* dataset requires multiple cameras (at least six) which remains a limiting factor. The following datasets rely more heavily on software as well as alternative sensors to relax this requirement and obtain reasonably approximate 3D ground truth in even more unrestricted settings.

The *3D Persons in the Wild* (*3DPW*) dataset (von Marcard *et al.*, 2018) is another recent dataset that has been adopted as a standard 3D pose benchmark. Here the authors rely on a single hand-held camera and inertial sensors (IMUs) to record sequences of one or two actors at a time in a variety of outdoor settings. IMUs suffer from a few issues such as measurement drift and lack of image synchronisation, but these are largely overcome through an optimisation scheme that combines the IMU readings, 2D keypoint detections, and a statistical body model which imposes anthropomorphic constraints. The accuracy of this scheme (2cm error) was verified against TotalCapture (Trumble *et al.*, 2017), an indoor 3D pose dataset that provides IMU readings, allowing *3DPW* to serve as a challenging benchmark for 3D pose estimation in the wild.

Eschewing the use of additional sensors entirely, the *Unite The People* (*UP-3D*) dataset (Lassner *et al.*, 2017) uses manual segmentation and keypoint annotations paired with an optimisation scheme (Bogo *et al.*, 2016) and human-in-the-loop verification to provide 3D ground truth for images from a variety of 2D pose estimation datasets: *LSP* (Johnson and Everingham, 2010), *LSP-extended* (Johnson and Everingham, 2011), *MPII HumanPose* (Andriluka *et al.*, 2014), and *FashionPose* (Dantone *et al.*, 2014). Fitting a 3D mesh to the images allows them to not only generate 3D pose ground truth but to also generate arbitrary 2D annotations transferred from the mesh that would be infeasible to collect manually. We make heavy use of this in Chapter 9. Interestingly, Lassner *et al.* (2017) also show that 2D keypoint detectors trained on these new annotations are more accurate and require less data than detectors trained on human-annotated 2D keypoints, owing to the spatial consistency of the former.

The *PedX* dataset (Kim *et al.*, 2019) is generated using a very similar approach, combining manual annotations and the same optimisation scheme as above, but with 3D data as an additional constraint. The dataset consists of a few video sequences recorded at street intersections using stereo cameras and lidar. Pixel-wise masks as well as 2D keypoints are labelled manually, and then a parametric body model is fit to the stereo-lidar data and annotations. This method is verified against a reference sequence recorded in a more controlled setting, and is shown to result in a small error (~2 cm, similar to *3DPW*).

In the same vein, Arnab *et al.* (2019) generate temporal 3D annotations for the large scale *Kinetics-600* action recognition dataset (Kay *et al.*, 2017). Unlike the *UP-3D* and *PedX* datasets however, they do not resort to manual 2D annotations given the scale of the underlying dataset, relying entirely on automatic methods. The resulting ground truth — while somewhat useful as an additional source of training data as they show — is thus significantly less precise. Is it more useful to generate smaller datasets with more precise annotations rather than larger ones with less precise ground truth? Evidence from the literature (e.g. Lassner *et al.* 2017) suggests that the former might be preferable, but it is not a resolved question.

Multi-person datasets

With the exception of *PedX*, all of the above datasets focus on one (or at most two in the case of *3DPW*) subjects at a time. However, there are other recent datasets devoted to the challenging multi-person setting.

The *MuPoTs-3D* dataset (Mehta *et al.*, 2018) is the multi-person follow-up to *MPI-INF-3DHP*. The test set consists of five indoor and 15 outdoor sequences with up to eight subjects, GT obtained from multiview marker-less mocap software. The corresponding training set (*MuCo-3DHP*) is obtained by compositing multiple subjects from *MPI-INF-3DHP* into single images in a depth-aware fashion.

The *Panoptic Studio* dataset (Joo *et al.*, 2019) is an in-studio dataset which focuses on social interactions between groups of people. Parsing such scenes requires the detection of subtle interaction cues. This requires very precise measurements that one could traditionally only obtain using marker-based system. Since the presence of visible markers might influence interactions between subjects, they instead resort to a special hardware setup consisting of hundreds of cameras — some with active sensors — distributed over a geodesic sphere. Measurements are integrated to obtain 3D motion ground truth for up to eight subjects per scene.

Synthetic datasets

As the preceding discussion makes clear, obtaining 3D pose ground truth paired with natural images is highly challenging. Only very recently have advances in data acquisition yielded data outside the traditional setting which focuses on a single actor per sequence inside a studio. Still, existing datasets are not without limitations, especially with regards to size, the number of subjects per scene, and the diversity of poses covered.

As a result, much effort has gone into generating synthetic data for 3D pose estimation. The benefits of synthetic data are limited by a domain gap in terms of appearance, but several works have shown that some improvement in performance can be gained by combining natural and synthetic images. The appearance gap also does not matter as much for pipeline approaches that first extract some intermediate representation from the image such as silhouettes before predicting 3D pose. Here, synthetic data has been shown to be especially useful.

Many older works have used synthetic data when 3D data was much harder to come by. Examples include: Agarwal and Triggs (2004), Shakhnarovich *et al.* (2003), Grauman *et al.* (2003), Sminchisescu *et al.* (2005), and Ionescu *et al.* (2009). More recent works include Chen *et al.* (2016), Ghezelghieh *et al.* (2016), Rogez and Schmid (2016), Varol *et al.* (2017), and Doersch and Zisserman (2019).

While many of the aforementioned works resort to synthetic data as a one-off means for training data generation or augmentation, Varol *et al.* (2017) put together the *SURREAL* dataset and benchmark which has found popular use, sometimes in modified form thanks to the accompanying open-source toolbox and access to the underlying data. They take 3D mocap sequences from the (CMU Graphics Lab Motion Capture Database) and apply the *MoSH* algorithm (Loper *et al.*, 2014) to the marker data to obtain *SMPL* model fits. These are then textured and rendered against images from the *LSUN* database (Yu *et al.*, 2015), which depict everyday indoor environments. As a by-product of the rendering process, they obtain semantic part labels, optical flow ground truth, as well as depth and normal information. They show that combining this data with natural data boosts part segmentation performance somewhat. One limitation of this data is that *SMPL* does not model variations in surface detail that can result for example from different hairstyles and loose-fitting clothing. This is addressed by Liang and Lin (2019), who present a dataset that includes renders with more realistic clothing in terms of surface details if not in terms of texture.

Chen *et al.* (2016) demonstrate that not just texture but also pose diversity is important for benefiting from synthetic training data. To this end, they learn a non-parametric, hierarchical model of the space of human poses that allows them to sample more diverse poses than are present in datasets such as the CMU Graphics Lab Motion Capture Database and *H36M* (Ionescu *et al.*, 2014). Each sample is used to pose a *SCAPE* model (Anguelov *et al.*, 2005). Clothing textures are deformed and added to the model which is rendered with random lighting and camera poses.

Synthetic data still suffers from a lack of realism in terms of appearance and Doersch and Zisserman (2019) attempt to sidestep this issue. They paste *SURREAL* models onto videos from the *Kinetics-400* data set (Kay *et al.*, 2017), and train a model on the resulting optical flow and keypoint motion as these suffer less from a realism gap. Indeed, they show that a model trained on this data outperforms a model trained on synthetic RGB data, even performing on par with a model trained on real RGB data.

Useful synthetic data can also be generated without rendering entirely, as shown by Rogez and Schmid (2016). Given 3D mocap data and a naturalistic dataset annotated with 2D keypoints they do the following: They project the 3D pose data, find a collection of images whose annotations each locally match the pose projection. These images are blended together to create a new image that corresponds to the original 3D pose. This is shown to be useful as an additional source of training data.

Most of the above datasets involve single persons rendered against simple backgrounds in geometry-free environments. Generating convincing looking synthetic people in a vacuum is hard enough. Positioning people plausibly in 3D scenes, potentially with multiple interacting people per scene brings its own set of challenges. Some recent work tackles this difficult problem setting, e.g. Hassan *et al.* (2021).

3.3 Methods

For the purpose of our discussion, it's useful to think of methods in this area as being primarily prediction-based or comparison-based. With the former (see Fig. 3.2a), a discriminative mapping is learned between image and 3D representation. Comparison-based methods in contrast recover pose and shape through some search procedure in which model-to-image comparison plays a central role. Two variants of these are: (i) model-based generative approaches (see Fig. 3.2b), where the parameters of some model are optimised to explain the image or some representation thereof, (ii) exemplar-based or dictionary-based methods, where an image representation is compared against a database or pose dictionary to retrieve the most likely pose.

This categorisation of course — like most abstractions — is an oversimplification. To the extent that this distinction was valid maybe one or two decades ago, the lines between different types of methods have only grown blurrier. For one, virtually all 3D human body recovery (3DHBR) methods nowadays rely on some discriminative component, not just prediction-based approaches.

Fully discriminative approaches based on convolutional neural networks (CNNs) have come to dominate this problem area. The methods define some target 3D representation, typically derived from 3D keypoints, and train a network to produce this representation given annotated data. While some methods directly map from image to 3D, e.g. Li and Chan (2014); Pavlakos *et al.* (2017), the majority take a so-called pipeline approach, which involves predicting some intermediate quantity — most commonly 2D keypoints. Such methods come in different flavours: Some use a separate lifting network that maps

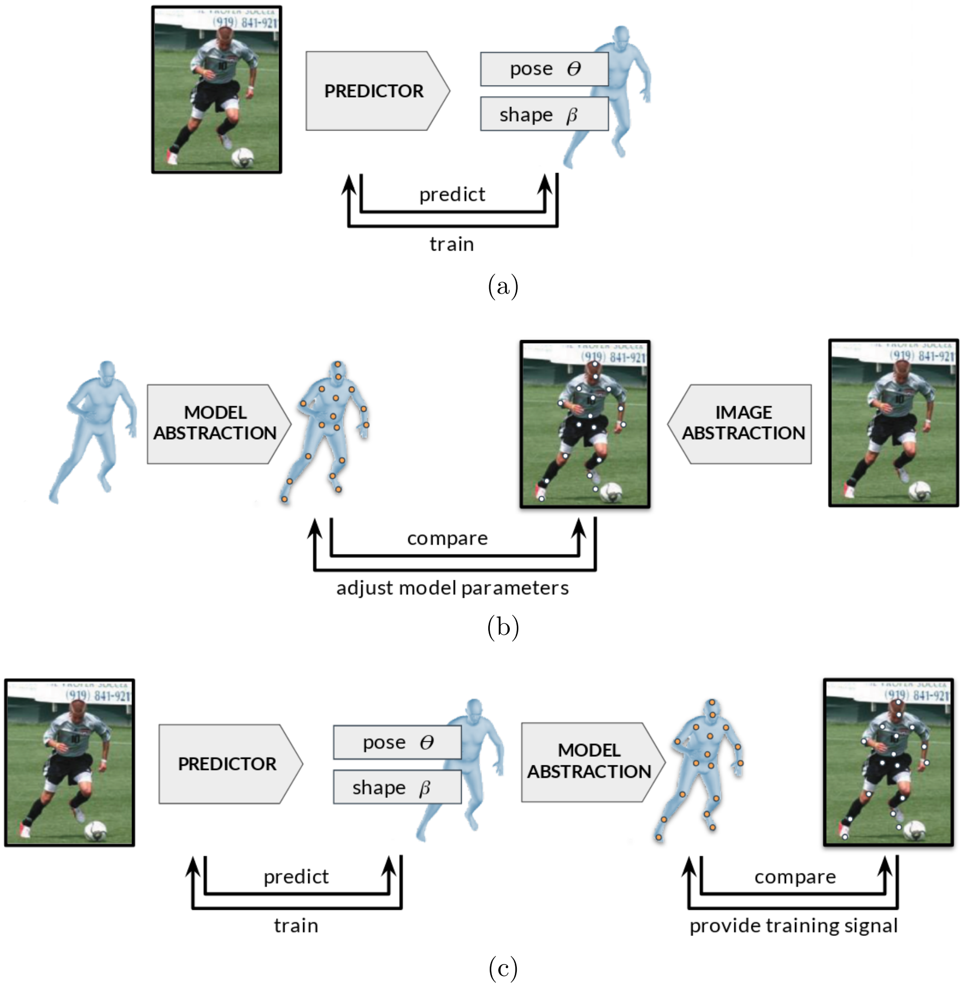


Figure 3.2: Schematic illustrations of representative classes of methods for 3D human body recovery methods. Prediction-based methods (a) rely on learning a mapping between an image and some representation of 3D pose and/or shape. Comparison-based methods rely on finding a common representation of the model and of the input image (e.g. keypoints) and updating model parameters to match the image observations (b), or using the input representation to query a database of pose exemplars. Recently, hybrid approaches (c) that combine elements of both are gaining popularity.

from 2D pose to 3D after abstracting away most image information, e.g. Moreno-Noguer (2017); Martinez *et al.* (2017). Others jointly train a network to predict 2D and 3D representations, e.g. Tekin *et al.* (2017); Habibie *et al.* (2019), fusing both image features and explicit 2D pose representations before the lifting step.

Modern comparison-based methods also almost exclusively resort to a pipeline approach. Some intermediate representation is extracted from the image in a discriminative manner in the first stage, e.g. 2D keypoints. These are then used as image evidence for model fitting (Zhou *et al.*, 2016b; Bogo *et al.*, 2016) or for a database look-up procedure (Chen and Ramanan, 2017). This was not the case with earlier model-based approaches, which relied on simpler and in many ways less reliable image abstractions such as low-level edges (Hogg, 1983), manually-annotated keypoints (Lee and Chen, 1985), or silhouettes obtained via background subtraction (Sminchisescu and Triggs, 2003).

Recently, more explicitly hybrid approaches have become popular (Tung *et al.*, 2017a; Kanazawa *et al.*, 2018; Pavlakos *et al.*, 2018b; Omran *et al.*, 2018) (see Fig. 3.2c), in which a human body model is embedded in a neural network and where the learning objective includes a term that compares model abstractions to an image. The target isn't merely the 3D pose representation; predictions of the model parameters are additionally encouraged to agree with some auxiliary observation such as 2D keypoints or silhouettes. Some methods even have an inner optimisation loop during training (Tomè *et al.*, 2017; Kolotouros *et al.*, 2019a), or additionally resort to test-time optimisation (Tung *et al.*, 2017a; Pavlakos *et al.*, 2018b; Zanfir *et al.*, 2020).

Given the above, rather than trying to organise methods into an artificial taxonomy, we will provide on overview of the literature centered around the following points:

- Given the important role that learning plays in 3DHBR methods, one aspect that distinguishes many methods is the 3D parametrisation of the human body that is chosen as the primary learning objective. Examples include 3D keypoint coordinates, various types of heatmaps, limb and joint rotations, body shape parameters, surface correspondences and pose embeddings.

- More so than with many computer vision problems, the notion of constraints plays an important role in recovering 3D human shape and pose. Outputs of 3DHBR systems often have to be additionally constrained in some way such that they represent anatomically valid poses, and this applies to both the training of discriminative methods as well as inference, especially if some search procedure is required. Such constraints can take on the form of local regularisers, e.g. to ensure limb symmetry or prevent interpenetration. Other methods rely on more global prior models of shape and pose, either with explicit probabilistic models or by baking this information into the output space itself. Some methods represent pose and/or shape using a weighted set of basis vectors, and the appropriate regulariser ensures that estimates don't stray away from known poses. Another important class of constraint are observation likelihoods, which measure the fidelity of the 3D estimates to some observation in the image.

- We then take an in depth look at the new crop of hybrid methods mentioned earlier, in which human body models play an important role during learning. Many methods in this area (including our own in Chapter 9) are pipeline approaches, relying on some intermediate representation prior to predicting body model parameters. The integration of a body model in a neural network also allows us to supervise the network with any quantity that can be derived from both the body model and the image. After discussing the various possibilities, we briefly turn to methods that handle temporal sequences of images.

- Finally, multi-person 3DHBR is starting to attract more interest and this setting brings with it additional challenges compared to the single-person setting.

3.3.1 Learning Objectives for Discriminative Methods

Coordinate regression

Many methods treat the task (at least partially if not fully) as a regression problem, i.e. by predicting the keypoint coordinates in metric space. Keypoint locations are predicted relative to one other, for example relative to some root joint (e.g. head or pelvis). Martinez *et al.* (2017) predict root-relative coordinates normalised by mean and standard deviation. They show that this is achievable with a very simple network operating purely on 2D keypoint coordinates from a separate method.

Some methods predict a keypoint's location relative to its parent along the kinematic tree, arguing that small, local displacements make for easier targets, e.g. Li and Chan (2014). Distances between neighbouring keypoints have smaller variances, and are more or less constant even for the same person. Left-right symmetry can also be exploited and information shared between different sub-predictions. The downside here however, is that errors can accumulate along the tree when reconstructing the full skeleton from individual predictions.

Other work suggests to predict denser offsets, i.e. to other joints besides either the parent or the root joint, e.g. Park *et al.* (2016); Mehta *et al.* (2017a). In the latter work it is argued that a combination of root-, parent- and grandparent-relative offsets are preferable, as the ideal offset will depend on the keypoint visibility and they show that this improves performance on hard poses in their setup (e.g. sitting).

A related representation is the distance matrix, proposed by Moreno-Noguer (2017), which encodes pair-wise distances between all joints, thus removing the location component from coordinate regression and only considering lengths. They argue that such a representation better captures similarity between related poses since encodes structural information more explicitly. It exhibits a higher correlation between 2D projection and 3D pose when both are represented with distance matrices as opposed to Cartesian coordinates. This representation is also invariant to global rotations, translations and scaling when normalised, thus obviating the need for pre-alignment of poses as is necessary for other methods. However, this necessitates an additional constrained optimisation step to recover the pose and resolve ambiguities.

The main difficulty with the regression approach in general is that the mapping between an image and the set of numerical values representing 3D keypoint locations is a highly non-linear one, and is not easy to learn. This is however mitigated by pipeline approaches, which predict intermediate quantities such as 2D keypoints, such as the simple and effective method of (Martinez *et al.*, 2017). Regression methods also tend to be more sensitive to image scale, requiring a tight crop around the person. These problems can be addressed by heatmap-based methods, where pose is either fully or partially represented in a way that corresponds to image pixels.

Volumetric heatmaps

An alternative to the direct regression approach is formulating the problem as one of pixel-wise prediction. This approach to a significant extent addresses the aforementioned difficulties with coordinate regression. In the context of 2D pose estimation, the choice between coordinate regression and pixel-wise prediction has been all but resolved in favour of the latter. See e.g. Tompson *et al.* (2014) for a discussion of this issue. In 2D, the correspondence between keypoints and pixel locations is the direct result of the annotation process. In 3D though, further processing is needed to establish this correspondence.

Pavlakos *et al.* (2017) extend the heatmap-based approach to 3D in a straightforward manner, namely with volumetric heatmaps. The final output of the network is a $64 \times 64 \times 64$ voxel grid per joint, x/y-axes correspond to pixel locations and the z-axis is a discretisation of the metric depth range [-1, 1]. A 3D Gaussian is placed at each joint in the grid. A stacked hourglass network (Newell *et al.*, 2016) is repurposed for the task. Instead of refining the same prediction from one module to the next, a coarse-to-fine approach is taken where the depth resolution is increased from 1 in the first module (i.e. 2D heatmap prediction) to 64 in four stages. The metric location of each joint is obtained by backprojecting the x/y values using a known camera calibration matrix and estimating the depth of the root joint based on the likely skeleton size and the size of its projection on the image. This approach is shown to significantly outperform a coordinate regression baseline, as well as competing pipeline approaches which up until that point were the better-performing ones. Heatmap-based approaches are more applicable to the multi-person case (Fabbri *et al.*, 2020) as there is no built-in assumption on the number of keypoints to expect, unlike with coordinate regression methods.

Another upside to such an approach is that heatmaps naturally capture uncertainty about keypoint locations, and are amenable to post-processing that takes this uncertainty into account. Direct coordinate regression in contrast does not admit straightforward reasoning about uncertainty. Heatmap outputs such as the above can for example be processed further using temporal filtering or with generative approaches that infer the most likely pose based on per-keypoint unary terms representing location uncertainty together with pairwise terms that explicitly capture structural constraints. Such generative approaches, most notably ones based on the pictorial structures framework (Felzenszwalb and Huttenlocher, 2005; Andriluka *et al.*, 2009) were at a time the dominant approach to 2D pose estimation but some have applied these to 3D pose with volumetric heatmaps, e.g. Kostrikov and Gall (2014); Kinauer *et al.* (2017).

While the aforementioned works predict x/y locations in pixel space — i.e. just the joints with visible projections, Sárándi *et al.* (2021) argue for volumetric heatmaps that represent metric space in all dimensions, not just in depth. This allows them to handle truncated people naturally, as the x/y predictions need not correspond to joints visible in the image crop. They demonstrate state-of-the-art results with a very simple network and heavy data augmentation, including artificially truncated and occluded examples. A differentiable layer that recovers the absolute root location using known focal length and 2D keypoint predictions in image space allows them to supervise their method with absolute coordinate locations as well.

Marginal heatmaps

The obvious downside of volumetric heatmaps is that despite requiring a lot of memory, only coarse discretisations of 3D space are possible given current hardware constraints. Kinauer *et al.* (2017) address this by resorting to coarse grids together with a per-voxel refinement offset. The most common way however to retain the benefits of heatmap-based representations while avoiding coarse outputs is to take a mixed approach, where 3D continuous predictions are tied to the 2D pixel grid in some manner. Another disadvantage of the parametrisations previously discussed, whether location coordinates or volumetric heatmaps is that they typically rely entirely on 3D ground truth.

To address these shortcomings, many methods resort to what can be referred to as marginal heatmaps. This involves decoupling the different spatial dimensions in some way, most commonly by predicting finely discretised 2D heatmaps (often corresponding to the pixel grid) corresponding to the x/y dimensions and then separately regressing to depth per 2D location Zhou *et al.* (2017) predict 2D heatmaps then regress separately to depth per pixel. Mehta *et al.* (2017b) resort to what they refer to as location maps, i.e. three 2D output maps, one per 3D coordinate. Their network regresses x/y/z values per pixel, and loss is only considered for pixels in the 2D vicinity of projected keypoints. At test time, the final values are read out from predicted 2D keypoint locations followed by kinematic skeleton-fitting.

An explicit decoupling of the z-dimension allows these methods to supervise their networks with a mixture of 2D and 3D annotations, which is especially useful given the availability of large-scale datasets with 2D keypoint annotations. This has shown to be very effective for increasing performance if not the main factor in some approaches. The experiments in Mehta *et al.* (2017a), which proposes a regression-based approach, demonstrate that the biggest effect on performance comes from transferring weights from a network trained to 2D pose. Li and Chan (2014) also demonstrate the benefits of pre-training on 2D data. More beneficial than pre-training is co-training with is facilitated by such representations (Tekin *et al.*, 2017; Sun *et al.*, 2017b).

One argument in favour of coordinate regression approaches is that they in principle allow for more precise outputs whereas heatmap-based approaches are limited in this regard due to the discretised spatial grid. Several approaches thus argue for the use of arg-softmax output layers, which combine advantages of heatmaps and numerical targets. Initially proposed in Chapelle and Wu (2010) for use in information retrieval, it has been since been rediscovered or used for robot learning Levine *et al.* (2016), feature detection Yi *et al.* (2016), 2D pose estimation Luvizon *et al.* (2018); Nibali *et al.* (2018) and 3D pose estimation Luvizon *et al.* (2018); Sun *et al.* (2018); Nibali *et al.* (2019). The idea is as follows: heatmaps necessitate the use of an argmax operation to get location coordinates. This tends to introduce quantisation errors due to the coarseness of network outputs. Instead, one can compute the expected location over heatmaps in different dimensions, and additionally supervise these with numerical targets. This still allows for the decoupling of dimensions, i.e. mixed 2D/3D training.

While Sun *et al.* (2018) conclude that 3D pose estimation performance does not necessarily benefit from such an approach, they show that lower resolution volumetric heatmaps degrade in performance more gracefully when combined with an arg-softmax-like loss function compared to using only the regular cross-entropy loss. They also include a nice ablation study that compares different prediction targets on an even footing. What is however not discussed in these works is that when the detector outputs two equally-confidence peaks in separate locations, these would be averaged out. An alternative would then be to perhaps predict spatial offsets at each location, as is often done for 2D pose estimation, e.g. in Insafutdinov *et al.* (2016). Sárándi *et al.* (2021) also confirm that arg-softmax is important for getting away with coarsely discretised volumetric predictions.

Joint rotations

Some methods parametrise 3D pose in terms of limb orientations. Luo *et al.* (2018) argue that the limitations of coordinate regression-based approaches can be overcome by predicting 2D keypoints together with normalised limb orientation vectors at pixels that correspond to limb projections. These are scale-independent, obviate the need for limb-length regularisation during training, and can in principle handle variable-sized inputs. As a post-processing step, a skeleton is recovered iteratively from the root joint outwards using: 2D joint locations, average limb orientation, known limb length ratios and scale information. Similarly, Liu *et al.* (2019a) and Xiang *et al.* (2019) both predict 2D keypoint locations and 3D unit vectors, but the former feeds this information to a 3D keypoint prediction network, whereas the latter fits a parametric human mesh.

Orientation vectors, however, only partially describe a joint's rotation, as these discard axial rotations. Some, e.g. Yoshiyasu *et al.* (2018), thus predict per-joint rotation matrices rather than orientation vectors, using a Gram-Schmidt orthogonalisation layer in the network to ensure valid outputs. They argue that this representation of rotations are more well-behaved compared to: (i) Euler angles which are discontinuous and (ii) quaternions which are invariant to sign flips. Additionally, they discretise the global rotation, treating it as a classification problem. A heuristic projection approach plus a learned decoder is used to obtain joint heatmaps.

Instead of using heuristic projections, learned decoders or optimisation to derive joint locations from joint rotations, these can be obtained deterministically by embedding a kinematic model into the network. Zhou *et al.* (2016a) for example use a kinematic skeletal model parametrised by global position, global orientation, and joint orientations. Bone lengths are assumed to be known and are not part of the estimate. They show that this improves over a baseline model which predicts joint locations directly. Contemporary approaches embed more sophisticated statistical models, such as *SMPL* (Loper *et al.*, 2015), that capture both skeletal pose and surface-level shape, besides also decoupling shape from pose entirely as described above. These methods resort to different rotation parametrisations as we will discuss shortly in Sec. 3.3.3.

Surface representations

There is evidence from the literature that richer output spaces can have a regularising effect and positively impact performance. Lassner *et al.* (2017) for example demonstrate this for 2D keypoint prediction, showing that training a network to predict 91 keypoints is better than using just 14 keypoints as a target on the same training data. Manually collecting such detailed annotations on in-the-wild images, let alone obtaining them in a consistent manner, is very difficult which is why Lassner et al. resort to a semi-automatic approach.

Some methods predict dense surface correspondences. One recent notable example is *DensePose* (Güler *et al.*, 2018), who find a way to get reliable manual annotations for surface correspondences to the *SMPL* body model, together with body part segmentations. While they do not recover 3D human pose or shape, subsequent work has shown this to be a useful proxy representation for lifting to 3D, e.g. Rong *et al.* (2019); Xu *et al.* (2019). Prior to the availability of such annotations for in-the-wild data, obtaining dense correspondence data for prediction and model fitting required depth data, e.g. Taylor *et al.* (2012); Pons-Moll *et al.* (2015).

Instead of predicting surface correspondences, Varol *et al.* (2018) predict volumetric 3D shape from data then fit a body model to that output. The model fitting step allows for recovery of fine detail, but voxel-based outputs suffer from the same limitations as volumetric heatmaps for 3D pose, e.g. a trade-off between resolution and memory requirements. Gabeur *et al.* (2019) propose an alternate non-parametric surface representation that sidesteps the resolution issue. They assume that the body surface can be split into hidden and visible depth maps, which applies to many frontal poses with limited self-occlusion. They then train network to predict both and use Poisson reconstruction to recover the surface from the resulting point cloud.

We discuss methods that directly recover the parameters of statistical body models in detail later, but as discussed above the surface is partly a function of a few pose-independent shape parameters, as well as the pose vector itself.which control the pose-dependent deformation of the body surface. Besides besides estimating pose, these methods typically just regress to the shape parameters.

Pose Embeddings

Some methods rely on learned pose embeddings that replace metric per-joint coordinates, arguing that poses occupy a lower-dimensional manifold in coordinate space. Such embeddings have long found use in 3D pose estimation, to make both discriminative mappings easier (Elgammal and Lee, 2004), as well as generative model-based optimisation (Sminchisescu and Jepson, 2004).

Some recent methods that use similar ideas includes Li *et al.* (2015), who train a network that maps image and pose vector to a common embedding space, such that the dot product of the two embedding vectors is highest when they match. At test-time, training set poses are scored against the input image and the average of the highest scoring poses is returned. Tekin *et al.* (2016) and Katircioglu *et al.* (2018) propose a more efficient method along similar lines. They train a denoising autoencoder for pose reconstruction and a separate network that maps images to the pose encoding space. At test-time, the image is mapped to the encoding space, and the decoder directly reconstructs the 3D pose.

Auxiliary tasks

Several works have shown that auxiliary learning targets can be useful, both when 3D pose annotations are and are not available. Ideas that have been explored in the literature include: binary depth relations, joint visibility and global orientation.

One idea is to use weak supervision in the form of relative orderings of joints, that is merely predicting whether or not a certain depth relation exists. The main motivation here is that it's easier to obtain such weak 3D supervision through manual annotation than it is to obtain precise 3D coordinates. Pons-Moll *et al.* (2014) thus propose to learn a set of 30 *PoseBits* that represent such relations as "joint x is in front of joint y". They additionally show that this is useful for sampling.

Pavlakos *et al.* (2018a) use a more exhaustive set of such "ordinal" relations, finding a way to get a minimal set that yields a global ordering on all 3D joints. They show that a network trained to predict 2D joints and ordinal relations using a ranking loss provides useful representations for a separate lifting stage, trained merely on unpaired 3D pose data. They also show that this helps in an end-to-end approach, with a network that is additionally trained to predict relative coordinates.

Wang *et al.* (2019) predict 2D joints, as well as the relative location of limb joints to torso (front, back, on-plane). In a second stage, they lift to 3D based on the preceding outputs together with image features. Additionally, they predict 3D pose in multiple stages: first from the torso outwards, then back towards the torso utilising previous predictions for each subsequent stage.

Another auxiliary task that helps with predictions, is e.g. predicting the person's orientation or more specifically their yaw angle relative to the ground plane, as in

Ghezelghieh *et al.* (2016). Unlike earlier work (e.g. Andriluka *et al.* (2010)) which used viewpoint prediction together with viewpoint-specific detectors and kinematic priors, here viewpoint prediction is used as a side task and source of information on the global configuration within the same network for the main objective. (Kiciroglu *et al.*, 2020)

Some methods show that joint visibility is a useful auxiliary task: Luvizon *et al.* (2018) show that training the network to predict joint visibility as a function of max-pooled heatmaps boost performance. As not all 3D datasets include joint visibility labels, Cheng *et al.* (2019) use a cylindrical person model to add occlusion labels, and show that these help when training 2D pose. They also train the lifting network to do the completion in 3D, but the biggest improvement results from training over sequences.

3.3.2 Constraints for Learning and Inference

There are different ways to constrain the outputs of a 3D human body recovery system. Broadly speaking, we can separate these into constraints that rely on prior knowledge and constraints based on observation likelihoods.

In the former case, a priori knowledge about the human body — whether its shape or the space of poses it can occupy — is used to guide learning and/or inference. This can include local kinematic constraints based on things like limb length statistics and joint angle limits, but also prior models of global pose probability. In some cases, some prior knowledge might be baked into the output space, e.g. with heatmap or dictionary-based representations. Monocular 3D pose estimation is an ill-posed task with ambiguities resulting from projection or occlusion. Accordingly, prior models of pose conditioned on partial observations also play a role but are less commonly used in modern methods, as these typically target point estimates of 3D pose.

The other class of constraints involves making sure that the output respects some observed quantity extracted from the image, such as 2D keypoints or silhouettes. Traditionally used by generative model-based approaches, discriminative methods increasingly add such terms to the loss function to provide a form of weak supervision.

Local kinematic constraints

The easiest, and thus most common way to constrain the output of pose estimation systems is through simple, local kinematic constraints. With discriminative methods, this takes the form of additional loss terms applied during training. In Zhou *et al.* (2017), bones in the same limb type (e.g. upper leg and lower leg) are required to have a constant ratio w.r.t. limbs in a canonical skeleton. Dabral *et al.* (2018) use the previous regulariser together with two other anatomically-inspired losses: They both penalise pairs of symmetric limbs that have unequal lengths, and penalise joints that bend unnaturally with heuristically set limits.

Similar penalty terms are also common in model-based based generative approaches to guide inference. Bogo *et al.* (2016) for example also use a heuristic penalty term that discourages unrealistic joint angles. They additionally use an interpenetration term enabled by their use of a surface rather than skeletal model. The *SMPL* mesh is approximated by a set of "capsules", which allows for fast and differentiable penalisation of self-intersection. A more precise version of this is proposed in Pavlakos *et al.* (2019a).

Global pose priors

The constraints previously discussed involve penalising unwanted configurations of individual limbs or at most pairs of limbs. Many methods instead resort to more global constraints in the form of prior models which assign probabilities to individual configurations of pose. However, among recent pose estimation methods, such explicit priors are not particularly common.

Rather than resort to fixed heuristics to penalise implausible poses, Akhter and Black (2015) point out that joint angle limits aren't static. To this end, they record a new mocap dataset with a wider range of poses and use it to learn a prior that distinguishes valid from invalid poses with pose-dependent joint angle limits, assigning uniform probability to valid poses.

Bogo *et al.* (2016) and Pavlakos *et al.* (2019a) use global priors that assign probabilities to individual poses: in the former case a Gaussian mixture model (GMM) and in the latter case a variational autoencoder that assumes a simple Gaussian latent space. While such formulations are convenient as well as effective provided the right amount of training data, they are not without flaws. Treating each pose vector as a point in global latent space neglects the compositional nature of pose, making it difficult to generalise to unseen poses (Lehrmann *et al.*, 2013). Jahangiri and Yuille (2017) also show that GMM-based priors reflect the statistics of the training set to a fault. Rare but not at all unusual poses (e.g. sitting poses) — rarity often being an artifact of a particular training set — are assigned low probabilities. This in all likelihood applies to VAE-based priors as well.

Several methods resort to implicit priors of pose by means of adversarial learning. A separate sub-network is trained to distinguish plausible from implausible poses and provide an error signal accordingly. Tung *et al.* (2017b) apply this approach to a latent PCA space of pose, while Kanazawa *et al.* (2018) train the discriminator directly on *SMPL* pose parameters. This biases the training away from implausible poses and allows them to use external mocap data unpaired with images. Kanazawa *et al.* (2018) additionally demonstrate that this allows for reasonable performance without paired annotations entirely. Yang *et al.* (2018) use the network from Zhou *et al.* (2017) as a generator. The discriminator receives 2D heatmaps, depth maps, image features and pairwise joint distance features and is applied to the generator outputs. This is shown to help with training the latter.

Drover *et al.* (2018) motivate their work with the observation that a predicted 3D pose might look plausible when projected from one viewpoint, but not from another. To this end, they train the lifting network from Martinez *et al.* (2017). Then, predictions are projected from a random view and passed to a discriminator which decides if these are plausible 2D configurations. This performs as well as a baseline trained with ground truth 2D joints. Similarly, in Wandt and Rosenhahn (2019) a network predicts 3D pose and camera parameters, and a separate GAN judges the plausibility of the pose. Chen *et al.* (2019) use a similar approach but with the express goal of only relying on 2D annotations to train their 3D pose estimation network. They also close the loop between 2D and 3D representations in more ways than one: For one, they apply a random 3D transformation, project the joints and use a discriminator to determine if 2D pose is a plausible one. Additionally, they lift transformed projection and compare to transformed 3D joints. They also invert the transformation, project the joints and compare to original 2D pose. Both of these works demonstrate strong performance compared to other weakly-supervised approaches.

There are two types of uncertainty when lifting 2D pose estimates to 3D: Uncertainty stemming from the depth ambiguity and uncertainty in the 2D observations themselves. Some methods have sought to explicitly take these into account.

Conditional prior models are also useful given the ambiguities involved in lifting 2D pose to 3D, e.g. depth ambiguities and uncertainty in the 2D observations themselves. Simo-Serra *et al.* (2012) focus on handling noisy 2D observations. Their idea is to generate several pose candidates from 2D joint detections. They achieve this by fitting Gaussians to observations in 2D and projecting these into 3D space. They then sample solution candidates in 2D, and efficiently select among these by minimising reprojection error to find the best pose.

Sharma *et al.* (2019) present a method that uses a sampling-based strategy, similarly to Simo-Serra *et al.* (2012), but to better handle the multi-modal nature of 3D lifting rather than to integrate 2D observation uncertainty. There, they predict 2D joints and ordinal relations. They train a conditional variational autoencoder (CVAE) Sohn *et al.* (2015) that can sample 3D poses from 2D and rank the samples based on agreement with the ordinal scores. The best-performing sample is picked as the final prediction. While this in principle allows for training with unpaired 3D data, they observe that the CVAE needs to see a pose distribution similar to the test set. Jahangiri and Yuille (2017) also propose a rejection sampling approach to obtain diverse 3D hypotheses conditioned on 2D pose estimates. They use the pose limit data from Akhter and Black (2015) to restrict samples to physically plausible poses while also accounting for observation uncertainty in 2D.

Constrained output spaces

An alternative to using prior models is to reparametrise the output space such that it reflects the statistics of pose. Volumetric heatmap-based representations for example by design impose an upper bound on the possible spatial dimensions of the output, informed by the statistics of skeleton dimensions (Pavlakos *et al.*, 2017), but these are very mild restrictions on predictions compared to the following methods.

Some methods take a classification-based approach to pose. Rogez *et al.* (2020) treat the full pose vector as an instance of one of 100 pose classes. For each person hypothesis, they propose to predict the pose class, as well as an offset to compensate for the difference between an instance and the class prototype. This pose prediction component is embedded in a network similar to *Faster R-CNN* (Ren *et al.*, 2015) which, in addition to predicting 3D pose, localises persons as well. Similarly, Yoshiyasu *et al.* (2018) discretise global pose into classes but use continuous regression for per-joint rotations Güler and Kokkinos (2019) in contrast take a classification approach to predicting individual joint rotations. This allows them to incorporate constraints on relative rotations as they only consider physically plausible joint extensions and leads to more robust predictions.

Exemplar-based approaches, e.g. Shakhnarovich *et al.* (2003); Mori and Malik (2006); Yasin *et al.* (2016); Chen and Ramanan (2017), take classification-based approaches to their logical extreme by relying on a database of poses. Chen and Ramanan (2017) argue that 3D human pose estimation boils down to ”2D pose estimation + matching“. Accordingly, the camera and 3D pose pair are retrieved from a database that best match 2D pose detections. The retrieved pose is warped with a simple procedure to better match the observed projection. Yasin *et al.* (2016) take a similar approach but with the k-nearest 3D poses. Exemplar-based methods are limited by the stored poses, but this arguably applies to CNN-based discriminative methods as well, which can be viewed as performing a sophisticated and efficient form of exemplar-matching against a database — the training set (Tatarchenko *et al.*, 2019).

Dictionary-based methods, e.g. Ramakrishna *et al.* (2012); Wang *et al.* (2014b); Zhou *et al.* (2016b); Tomè *et al.* (2017), represent the space of poses as an overcomplete dictionary of basis vectors. This can be accomplished for example by applying PCA separately to individual actions (Ramakrishna *et al.*, 2012) or to pose clusters (Tomè *et al.*, 2017) and concatenating the resulting basis vectors. Together with the appropriate regulariser on the basis vector weights, e.g. one that encourages sparsity (Wang *et al.*, 2014b), this ensures that outputs somewhat reflect the statistics of pose. As pointed out in Akhter and Black (2015) however, this does not guarantee outputs that are plausible — provided we assume all bones are intact.

Observation likelihoods

So far, we have mostly discussed constraints based on prior knowledge. These are independent of the underlying image and are derived from our knowledge of the human body, e.g. its common shapes and poses. Another very important type of constraint — albeit one that has become less important given the rise of discriminative methods — are those derived from image observations, e.g. edges, silhouettes, keypoints and depth information. Prior to the availability of large annotated datasets and strong discriminative methods, 3DHBR methods typically relied on optimisation schemes where a model of the human body was matched against simple image abstractions. These were often difficult to extract reliably and also difficult to match the model against, e.g. low-level edges (Hogg, 1983) and binary silhouettes obtained with background subtraction (Sminchisescu and Triggs, 2003; Balan *et al.*, 2007c) or more involved methods (Guan *et al.*, 2009; Gall *et al.*, 2010). In contrast, 2D keypoints are significantly easier to recover in a differentiable manner from human body models. While some early approaches relied on manually-annotated keypoints (Lee and Chen, 1985; Taylor, 2000; Guan *et al.*, 2009), advances in 2D pose estimation have made it easy to also predict 2D keypoints automatically and reliably from images. 2D keypoints are accordingly a very popular choice of constraint both during training (Kanazawa *et al.*, 2018) and inference (Bogo *et al.*, 2016). Constraints based on image observations are very relevant to the hybrid methods we present next, and will thus continue the discussion thereof in the next section.

3.3.3 Hybrid Methods for Mesh Recovery

In this section we focus on a class of methods that is particularly relevant to this thesis, as we present one such method in Chapter 9. These are methods that combine elements of prediction-based and comparison-based approaches in a fairly novel way that has been enabled by advances in data acquisition, parametric body modelling, neural network training and differentiable rendering. These approaches address one of the key shortcomings of comparison-based approaches — especially generative, model-based approaches — namely initialising the search for good model parameters. The integrated model also helps to constrain the predictions of the discriminative prediction function during training and enables more flexible supervision than would otherwise be possible.

The basic structure of such approaches is as follows: A discriminatively-trained *encoder* predicts the parameters of a *body model*, possibly after predicting an intermediate representation (e.g. keypoints or silhouettes). Many recent works integrate the *SMPL* model (Loper *et al.*, 2015) as it produces a mesh from pose and shape representations with differentiable, as well as mostly linear operations. The body model can thus be instantiated from network shape and pose predictions, and is then followed by a *decoder* that renders the resulting skeleton and mesh in some form, e.g. by projecting skeleton keypoints to 2D or by rendering the mesh surface. The full model can be trained end-to-end using a variety of losses on both the model parameters, but also on any quantity that can be derived from the model. While this not common, in principle the presence of the model allows for iterative optimisation at test-time to refine the parameters on the basis of the initial prediction (Tung *et al.*, 2017a; Pavlakos *et al.*, 2018b; Zanfir *et al.*, 2020).

In the following, we will focus on different aspects relevant to such methods: (i) intermediate representations prior to predicting body model parameters, (ii) which body models are used and how they can be mapped back to the image, (iii) how such approaches can be supervised, and (iv) extensions of such models to image sequences.

Intermediate Representations

One of the main aspects that distinguishes methods in this space is the type of proxy representation used prior to the lifting step. Various proxies have been explored in the literature including in 2D (e.g. keypoints, segmentations) and 3D (e.g. keypoints, mesh vertex locations, limb orientations). Another line of work involves breaking down the global shape and pose representation into a hierarchy of part-based representations.

Some methods avoid proxy representations altogether (Kanazawa *et al.*, 2018), but it has been shown that intermediate representations can result in more sample-efficient learning as well as better reconstruction results (Chapter 9). Examples include 2D keypoints & silhouettes (Pavlakos *et al.*, 2018b)), body part segmentations (Chapter 9), surface correspondences (Rong *et al.*, 2019; Xu *et al.*, 2019; Zhang *et al.*, 2019) or even meshes (Kolotouros *et al.*, 2019b) and voxel reconstructions (Varol *et al.*, 2018). In contrast to our work, Rong *et al.* (2019) find that the benefits of using intermediate representations are limited to non-existent depending on how the network is supervised. We're not sure what explains this discrepancy, but in any case there are other benefits to using intermediate representations, such as the use of synthetic data or using the proxy representation for self-supervision of the lifting network. Pavlakos *et al.* (2018b) for example separately train networks that map from silhouette and 2D keypoints to *SMPL* shape and pose parameters. These sub-networks are then embedded into an end-to-end pipeline. Another benefit of intermediate representations is improved interpretability, as body model parameters are global vectors less tied to the image than 2D pixel-wise representations.

The aforementioned methods treat the pose parameters as a global representation to be estimated in one shot, but others break it down into components. (Güler and Kokkinos, 2019) predict individual joint rotations separately by pooling information from connected joints. For this, they use 2D keypoint estimates. (Zhang *et al.*, 2019) also use 2D keypoint locations as a guide. In addition to using a global IUV map as an intermediate representation, they use *RoI-pooling* (Ren *et al.*, 2015) to extract separate per-keypoint features. From these they predict part-level IUV maps and intermediate rotation features before estimating the full global pose. A dropout-like mechanism (Srivastava *et al.*, 2014) at the level of parts is used to encourage robustness against bottom-up errors and occlusions. Georgakis *et al.* (2020) also exploit the hierarchy of the kinematic chain to avoid the standard "features-in-parameters-out“ approach as they call it. The *SMPL* skeleton is split into six kinematic sub-chains including a root chain that serves as parent to the rest. The pose parameters of all chains are estimated separately in an iterative fashion, and child chains additionally depend on the root estimate. They show that this results in more graceful degradation under occlusion compared to competing approaches.

While some methods use surface correspondences to the *SMPL* mesh as a proxy representation, others output the surface itself in 3D. Varol *et al.* (2018) output a voxel representation of the *SMPL* mesh after first estimating 2D keypoints, 3D keypoints and part segmentations. *SMPL* can be subsequently fit to the volumetric output which is limited in terms of resolution. Kolotouros *et al.* (2019b) take a more structured approach and predict the vertex locations of the *SMPL* mesh. They use graph convolutions that smooth predictions within local neighbourhoods while taking image features into account. A simple regressor predicts the *SMPL* shape and pose parameters using the mesh as input. The intermediate output thus captures some surface details (e.g. clothing, hair) as a by-product, but Zheng *et al.* (2019) show that these are more easily extracted by first estimating *SMPL* parameters (i.e. posing the model) and then refining the corresponding mesh with surface normal estimates. Choi *et al.* (2020) similar to Kolotouros *et al.* (2019b). use graph convolutions to predict a mesh but instead take a coarse-to-fine approach. After lifting 2D pose estimates to 3D using the *SimpleBaseline* network (Martinez *et al.*, 2017), they predict a sparsified human mesh consisting of 96 vertices, that is refined in multiple steps until the full *SMPL* mesh is produced. Moon and Lee (2020) propose a so-called lixel representation for 3D coordinates. They output three 1D heatmaps per joint or per vertex (i.e. one per spatial coordinate). This is a more efficient representation than volumetric heatmaps (e.g. (Pavlakos *et al.*, 2017)) and that scales to the *SMPL* mesh. Vertex coordinates are thus predicted in a way that is registered spatially to the image along each dimension separately, and such that prediction uncertainty is represented.

Body Models

Most methods in this space use *SMPL* as the parametric body model but there are some exceptions. Xu *et al.* (2019) use a custom, *SMPL* -like model whose parameters are learned from all scans in the *CAESAR* database (Robinette and Daanen, 1999). The goal is to learn a richer, gender-neutral shape space. They also add 28 joints to the default *SMPL* skeleton to capture fingers and five additional joints for spine and head. Xiang *et al.* (2019) use a model derived from *SMPL* — *Frankenstein* (Joo *et al.*, 2018) — that additionally captures facial expressions besides fine-grained hand pose, but does away with the pose blend shapes. The latter leads to a drop in realism with respect to vanilla *SMPL* but it's not clear if this provides benefits in terms of ease of optimisation. Zanfir *et al.* (2020)

As the mapping from parametric body model to keypoints or part segmentation maps is deterministic, decoders are typically parameter-free. Keypoints can be recovered with simple projection and binary silhouettes can either be obtained via projecting mesh points (Pavlakos *et al.*, 2019b) or via differentiable rendering (Loper and Black, 2014; Henderson and Ferrari, 2018). In the latter case, we can also obtain more complicated more abstractions such as part segmentations provided that labels are attached to mesh vertices.

Deriving keypoints or silhouettes necessitates estimating camera parameters together with the *SMPL* parameters. Most methods apply a weak perspective model (or scaled orthographic projection), which requires estimating just three parameters (Kanazawa *et al.*, 2018). This approximation works because humans tend to be compact in the depth dimension relative to their distance to the camera.

One exception to the use of parameter-free decoders is the method of Tan *et al.* (2017), who resort to a learned decoder instead. Theirs is a neural network trained to reconstruct silhouettes from *SMPL* parameters on an artificial dataset. This decoder is then kept fixed, and the encoder is trained to produce parameters that result in accurate silhouettes. The encoder can additionally be trained using a regression loss on the model parameters if available.

Types of Supervision

The embedded parametric body model enables a diverse array of losses for training these methods. The most common ones are losses on body model parameters, mesh vertex locations and 3D and/or 2D keypoints. Some methods also use part segmentations, dense surface correspondences, optical flow and photometric losses as well.

Most methods use direct supervision in the form of body model parameters that encode shape and pose whenever available. Kanazawa *et al.* (2018) among others uses the default axis-angle rotation used to represent poses in *SMPL* , but others argue for the use of rotation matrices as these are better behaved and can lead to faster convergence if not to better results (Lassner *et al.*, 2017; Pavlakos *et al.*, 2018b; Omran *et al.*, 2018; Kolotouros *et al.*, 2019b). Similarly, Zhou *et al.* (2019b) propose an alternate 6D-representation for rotations that — unlike axis-angle representations, Euler angles, and quaternions — is also continuous, and leads to better empirical results on different pose estimation tasks than full rotation matrices. Kolotouros *et al.* (2019a) use this representation for 3D shape and pose estimation. Ground truth body model parameters are difficult to obtain, but Omran *et al.* (2018) and Rong *et al.* (2019) show that a small amount of such annotations is sufficient when combined with alternate sources of supervision, such as 3D or 2D keypoints. We will discuss other methods that attempt to do away with such supervision entirely. (Kanazawa *et al.*, 2018) additionally show good results with indirect supervision using body model parameters from a motion capture dataset. An adversary is trained on this data to judge the results of the main regressor, and in the ”unpaired setting“ the regressor is trained to fool the adversary with only 2D keypoint supervision.

Besides supervision through body model parameters, the most common type is supervision via keypoint losses. 2D keypoint annotations are available in abundance for in-the-wild datasets and help with generalising to in-the-wild settings. 3D keypoints can be used for supervision as well, either derived from the body model parameters or from 3D pose datasets such as *H36M* (Ionescu *et al.*, 2014) or *MPI-INF-3DHP* (Mehta *et al.*, 2017a). This requires a separate regressor from the *SMPL* mesh to the keypoints. Similarly, some methods use mesh vertex error as a loss as this is richer than merely keypoint supervision. Only training with a loss on the keypoints can result in the prediction of unnatural shapes. (Güler and Kokkinos, 2019) demonstrate that a combination of multiple losses is important for achieving balanced results that improve both shape and pose estimation.

Zanfir *et al.* (2020) use a loss based on semantic part segmentation, similar to the one used by Zanfir *et al.* (2018a). This loss doesn't use differentiable rendering, but simply projects all mesh vertices to the image. Each pixel of a bottom-up part segmentation attracts the closest mesh vertex with the same part label. This can be viewed as ICP-like process with 2D-3D correspondences.

Rong *et al.* (2019) look into question of which data to acquire for supervision, as getting *SMPL* parameters in-the-wild is tricky *DensePose* annotations are a good proxy for supervision (much more so than sparse 2D keypoints). Xu *et al.* (2019), besides arguing for predicting the body pose parameters from an IUV map (obtained from *DensePose*), use a differentiable rasterizer that allows them to supervise the network with a combination of losses from part maps / 2D pose / IUV map and on the parameters. Additionally, they produce a new synthetic dataset (*MOCA*) with 2M images which gives an edge in performance.

While most recent methods rely on purely semantic correspondences to the image as an optimisation target, some additionally use low-level information as optimisation targets — much more common in older methods. Besides using keypoints and silhouettes for supervision, Tung *et al.* (2017a) also use motion information as source of supervision during training but also at test-time to refine body parameter estimates. Given model predictions in successive frames, vertex displacements in time are compared against estimates of optical flow and the discrepancy is used as a loss signal. Similarly, direct photometric losses are uncommon in this space given the diversity of human appearance that is not captured by statistical body models. These are more commonly used for hand pose estimation and 6DOF rigid object reconstruction and tracking when the object is known beforehand. Recent methods that leverage these for 3D human shape and pose estimation are those of Pavlakos *et al.* (2019b) and Rueegg *et al.* (2020).

One way to use such appearance-based losses is together with multi-view images or monocular video sequences.Pavlakos *et al.* (2019b) train their model with batches consisting of images of the same person. They estimate model parameters for each image, figure out which vertices are visible across images and project these to get the corresponding texture value. Colour consistency among corresponding vertices is then used as a supervision signal. At test-time their method is applied to single images, and they find that this training procedure helps a little both in the unpaired setting as well as when 3D ground truth is available. This also allows them to use unlabelled in-the-wild video for training but this does not appear to improve in-the-wild performance.

All of the above methods use 3D supervision in some form or other. Most use at least some amount of direct supervision, but can use unpaired 3D data to constrain network outputs such as *HMR*. In contrast, Rueegg *et al.* (2020) attempt to tackle the challenging problem of learning 3D without access to any 3D data. They use a *CycleGAN*-like architecture (Zhu *et al.*, 2017a), where one loop learns to go back and forth between an image to separate part segmentation, appearance features, and the background. The second loop connects the part segmentations to *SMPL* body parameters. The network is supervised by objective that encourages the estimated *SMPL* parameters together with the background, appearance features, and part segmentations to produce an image that matches the input.

Image Sequences

Several works extend these methods to video, e.g. Sun *et al.* (2019), Kanazawa *et al.* (2019) and Kocabas *et al.* (2020). The latter work is an extension of *SPIN* to video, while Kanazawa *et al.* (2019) extend *HMR* to receive multiple frames. These are jointly encoded, *SMPL* parameters are predicted both at the current frame and at a temporal offset in both directions. A separate single-frame model is trained to mimic this joint frame encoding, allowing for the hallucination of motion from a single frame. They show that this also helps improve single-frame 3D shape and pose estimation performance.

3.3.4 Multi-person 3D Pose Estimation

As with multi-person 2D pose estimation, there are two classes of methods that address the task in 3D: person-first or top-down methods which start from person detections and for each estimate 3D person, and keypoint-first or bottom-up methods which first proceed in a person-agnostic manner and then group detected keypoints into person hypotheses. However, unlike with 2D pose estimation, top-down methods here often resort to an extra optimisation step to ensure global consistency of some quantity such as distance to the observer or camera parameters.

Person-first methods

In a previous section, we described the method of Rogez *et al.* (2020). Others include the work of Zanfir *et al.* (2018a), in which they first detect people in the scene, then use the multi-task network from Popa *et al.* (2017) to produce 2D/3D joints as well as semantic labels. Additionally, an initial shape and pose estimate is made per person per-frame. They then optimise over all people in the scene with terms that consider ground plane constraints as well as a penalty term for simultaneous volume occupancy.

Moon *et al.* (2019) first detect people then for each person they estimate the absolute depth of the root joint as well as root-relative joint offsets. Dabral *et al.* (2019) replace the keypoint head in Mask R-CNN (He *et al.*, 2017) with a Stacked Hourglass Network (Newell *et al.*, 2016). Each 2D estimate is with Simple 3D Baseline network (Martinez *et al.*, 2017). Afterwards, they optimise for global positions and focal length.

Keypoint-first methods

Methods that produce bottom-up 3D-pose evidence for multiple people at once includes the work of Mehta *et al.* (2018). They build on the location maps proposed in Mehta *et al.* (2017b) which were designed for single person 3D pose estimation, meaning that they assume all joints are visible and that at most one instance per joint needs to be read out from the corresponding map. This is not the case with multiple people, so they propose so-called occlusion-robust pose maps. The key idea is to introduce redundancy such that information on different joints can be obtained in multiple ways. The full-pose can be read out at torso locations, per-joint poses can be read out at the corresponding 2D locations, and complete limb poses are encoding at any 2D location of one of the limb joints. An occlusion-aware inference strategy reads out 3D information as needed.

Zanfir *et al.* (2018b) predict 2D joint locations and use a learned scoring function to group these into limbs. An integer linear optimisation is performed to assemble skeletons that respect kinematic constraints. The same network that predicts 2D joints also trained to predict full 3D pose at every pixel corresponding to a limb as a source of additional evidence.

The problem with approaches that proceed from the bottom-up is the potential for decoding conflicts when estimates for similar joints belonging to different people coincide or are in close proximity. Mehta *et al.* (2020) propose to address this as follows: A fully-convolutional network predicts 2D joint maps, 2D part affinity fields, and parent-relative 3D offsets. Rather than assemble 3D pose based on bottom-up evidence, this is fed to a separate network that predicts the full 3D pose for each person hypothesis.

3.4 Summary

In this chapter we reviewed recent work on 3D human shape and pose estimation. We discussed different variants of this task, as it is less well-defined than pedestrian detection. This included a discussion of various 3D representations of the human body, in particular the class statistical mesh models relevant to our work. We then described relevant datasets as well as the unique challenges of obtaining 3D ground truth especially outside of controlled studio settings.

We then dedicate the rest of the chapter to reviewing methods. We start with discriminative approaches and specifically the different kinds of targets used to train such methods, as these are relevant to recent methods with a strong generative component. The notion of constraints plays an important role in this area, whether to constrain predictions of discriminative methods or to constrain the search for suitable parameters in generative approaches. There are many types of constraints that include simple physical or anatomical constraints and probabilistic priors of pose, but also output spaces that incorporate constraints from the outset. While these are constraints that rely on prior knowledge, another important class of constraints are ones based on image observations such as keypoints and silhouettes. Before concluding, we review hybrid methods that integrate statistical body models into a discriminative NN-based predictor. Finally, we briefly discuss methods designed for handling multiple people, which are starting to attract more interest.

Part I

DETECTION

4

Lessons from a Decade of Pedestrian Detection: 2004-2014

PAPER-BY-PAPER results make it easy to miss the forest for the trees. We analyse the remarkable progress of the decade from 2004–2014 by discussing the main ideas explored in the 40+ detectors present in the *Caltech Pedestrian Dataset* (*Caltech*) (Dollár *et al.*, 2009b). We observe that there exist three families of approaches, all reaching similar detection quality. Based on our analysis, we study the complementarity of the most promising ideas by combining multiple published strategies. This new decision forest detector achieves the best performance on the challenging *Caltech* dataset in July 2014.

This work has been published at the "Computer Vision for Road Scene Understanding and Autonomous Driving" workshop (Benenson *et al.*, 2014). Rodrigo Benenson was the lead author, Mohamed Omran conducted most of the experiments, and Jan Hosang contributed experiments incorporating context using *2Ped* (Ouyang and Wang, 2013b) in Sec. 4.4.2, plots, writing, and analyses.

4.1 Introduction

The aim of this chapter is to review progress over a decade of pedestrian detection between 2004 and 2014 (40+ methods), identify the main ideas explored, and try to

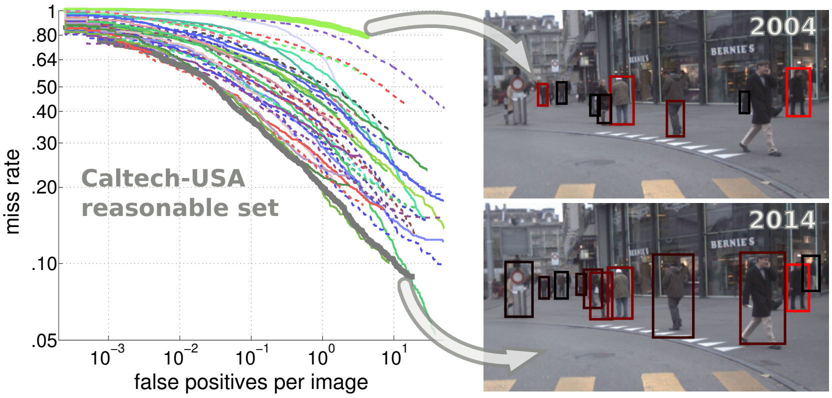
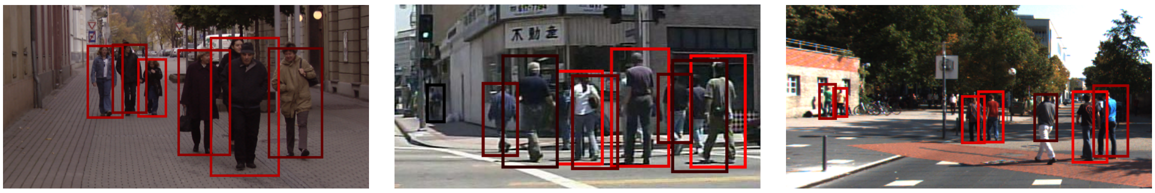


Figure 4.1: The last decade has shown tremendous progress on pedestrian detection. What have we learned out of the 40+ proposed methods?



(a) *INRIA* test set (b) *Caltech* test set (c) *KITTI* test set

Figure 4.2: Example detections of a top performing method (*SquaresChnFtrs*).

quantify which ideas had the most impact on final detection quality. In the next sections we review existing datasets (Sec. 4.2), provide a discussion of the different approaches (Sec. 4.3), and experiments reproducing/quantifying the recent years' progress (Sec. 4.4, presenting experiments over ~ 20 newly trained detector models). Although we do not aim to introduce a novel technique, by putting together existing methods we report best detection results at the time of publication on the challenging *Caltech* dataset.

4.2 Datasets

In the period covered by this analysis, multiple public pedestrian datasets have been collected over the years: *INRIA* (Dalal and Triggs, 2005), *ETH* (Ess *et al.*, 2008), *TUD-Brussels* (Wojek *et al.*, 2009), *Daimler* (Enzweiler and Gavrila, 2009) and the related *Daimler Stereo* (Keller *et al.*, 2009), *Caltech* (Dollár *et al.*, 2009b), and *KITTI* (Geiger *et al.*, 2012) are the most commonly used ones. They all have different characteristics, weaknesses, and strengths.

INRIA is amongst the oldest and as such has comparatively few images. It benefits however from high quality annotations of pedestrians in diverse settings (city, beach, mountains, etc.), which is why it was commonly selected for training (see also Sec. 4.4.4). *ETH* and *TUD-Brussels* are mid-sized video datasets. *Daimler* is not considered by all methods because it consists of greyscale images. *Daimler Stereo*, *ETH*, and *KITTI* provide stereo information. All datasets except *INRIA* are derived from video sequences, and thus support the use of optical flow as an additional cue.

At the time we conducted this analysis, *Caltech* and *KITTI* were the predominant benchmarks for pedestrian detection. Both are comparatively large and challenging. *Caltech* stands out for the large number of methods that have been evaluated side-by-side. *KITTI* stands out because its test set is slightly more diverse, but had not been yet used as frequently as Caltech. *INRIA*, *ETH* (monocular), *TUD-Brussels*, *Daimler* (monocular), and *Caltech* are available under a unified evaluation toolbox. *KITTI* uses its own separate evaluation server with unpublished test annotations. Both benchmarks maintain an online ranking which provide a quick overview of results.

In this chapter we primarily use *Caltech* for comparing methods, and *INRIA* and *KITTI* as secondary benchmarks. See Fig. 4.2 for example images. *Caltech* and *INRIA* results are measured in log-average miss rate (laMR, lower is better), while *KITTI* uses area under the precision-recall curve (AUC, higher is better). Chapter 2 contains a more comprehensive discussion of datasets (Sec. 2.2) as well as of these evaluation metrics (Sec. 2.1).

Individual papers usually only show a narrow view over the state of the art on a dataset. Having an official benchmark that collects detections from all methods greatly facilitates comparisons against the state of the art for both authors and reviewers. The collection of results enable retrospective analyses such as the one we present next.

4.3 Elements of Pedestrian Detectors

Tab. 4.1 and Fig. 4.3 together provide a quantitative and qualitative overview over 40+ methods whose results are published on the official *Caltech* benchmark up until July 2014. Methods marked in italics are our newly trained models (described in Sec. 4.4). We refer to all methods using their *Caltech* benchmark shorthand. Instead of discussing the methods' individual particularities, we identify the key aspects that distinguish each method (ticks of Tab. 4.1) and group them accordingly. We discuss these aspects in the next subsections.

Brief Chronology. Our analysis starts with the seminal *Viola-Jones* (*VJ*) detector, which Viola *et al.* (2003) applied to the task of pedestrian detection after its success with face detection (Viola and Jones, 2001). Soon after, Dalal and Triggs (2005) introduced the landmark *HOG* detector, which later served as a building block for the now classic *Deformable Part Model* detector (*DPM* — or *LatSvm* on the *Caltech* leaderboard) (Felzenszwalb *et al.*, 2008). In 2009, *Caltech* was introduced (Dollár *et al.*, 2009b) with a comprehensive quantitative comparison of seven pedestrian detectors. Since then, the evaluation metric changed from per-window (FPPW) to per-image (FPPI), once the flaws of the per-window evaluation were identified (Sec. 2.1, Dollár *et al.* 2012b). This new metric was more suited for evaluating detection rather than classification performance, and weaknesses in older methods that had otherwise been obscured came to light.

About one third of the methods considered here were published during 2013, reflecting a renewed interest in the problem. Similarly, half of the *KITTI* results for pedestrian detection were submitted in 2014.

4.3.1 Training Data

Fig. 4.3 shows that differences in detection performance are, unsurprisingly, dominated by the choice of training data. Methods directly trained on *Caltech* systematically perform better than methods that attempt to generalise from *INRIA* to *Caltech*. Tab. 4.1 gives

Method	MR	Family	Features	Classifier	Context	Deep	Parts	M-Scales	More data	*Feat. type*	*Training*
VJ (Viola and Jones, 2004)	94.73%	DF	✓	✓						Haar	I
Shapelet (Sabzmeydani and Mori, 2007)	91.37%	-	✓							Gradients	I
PoseInv (Lin and Davis, 2008)	86.32%	-					✓			HOG	I+
LatSvm-V1 (Felzenszwalb *et al.*, 2008)	79.78%	DPM					✓			HOG	P
ConvNet (Sermanet *et al.*, 2013)	77.20%	DN				✓				Pixels	I
FtrMine (Dollár *et al.*, 2007)	74.42%	DF	✓							HOG+Color	I
HikSvm (Maji *et al.*, 2008)	73.39%	-		✓						HOG	I
HOG (Dalal and Triggs, 2005)	68.46%	-	✓	✓						HOG	I
MultiFtr (Wojek and Schiele, 2008)	68.26%	DF	✓	✓						HOG+Haar	I
HogLbp (Wang *et al.*, 2009)	67.77%	-	✓							HOG+LBP	I
AFS+Geo (Levi *et al.*, 2013)	66.76%	-			✓					Custom	I
AFS (Levi *et al.*, 2013)	65.38%	-								Custom	I
LatSvm-V2 (Felzenszwalb *et al.*, 2010)	63.26%	DPM		✓			✓			HOG	I
Pls (Schwartz *et al.*, 2009)	62.10%	-	✓	✓						Custom	I
MLS (Nam *et al.*, 2011)	61.03%	DF	✓							HOG	I
MultiFtr+CSS (Walk *et al.*, 2010)	60.89%	DF	✓							Many	T
FeatSynth (Bar-Hillel *et al.*, 2010)	60.16%	-	✓	✓						Custom	I
pAUCBoost (Paisitkriangkrai *et al.*, 2013)	59.66%	DF	✓	✓						HOG+COV	I
FPDW (Dollár *et al.*, 2010)	57.40%	DF								HOG+LUV	I
ChnFtrs (Dollár *et al.*, 2009a)	56.34%	DF	✓	✓						HOG+LUV	I
CrossTalk (Dollár *et al.*, 2012a)	53.88%	DF			✓					HOG+LUV	I
DBN−Isol (Ouyang and Wang, 2012)	53.14%	DN					✓			HOG	I
ACF (Dollár *et al.*, 2014)	51.36%	DF	✓							HOG+LUV	I
RandForest (Marín *et al.*, 2013)	51.17%	DF		✓						HOG+LBP	I&C
MultiFtr+Motion (Walk *et al.*, 2010)	50.88%	DF	✓						✓	Many+Flow	T
SquaresChnFtrs (Benenson *et al.*, 2013)	50.17%	DF	✓							HOG+LUV	I
Franken (Mathias *et al.*, 2013)	48.68%	DF		✓						HOG+LUV	I
MultiResC (Park *et al.*, 2010)	48.45%	DPM			✓		✓	✓		HOG	C
Roerei (Benenson *et al.*, 2013)	48.35%	DF	✓					✓		HOG+LUV	I
DBN−Mut (Ouyang *et al.*, 2013)	48.22%	DN			✓		✓			HOG	C
MF+Motion+2Ped (Ouyang and Wang, 2013b)	46.44%	DF			✓				✓	Many+Flow	I+
MOCO (Chen *et al.*, 2013)	45.53%	-	✓		✓					HOG+LBP	C
MultiSDP (Zeng *et al.*, 2013)	45.39%	DN	✓		✓	✓				HOG+CSS	C
ACF-Caltech (Dollár *et al.*, 2014)	44.22%	DF	✓							HOG+LUV	C
MultiResC+2Ped (Ouyang and Wang, 2013b)	43.42%	DPM			✓		✓	✓		HOG	C+
WordChannels (Costea and Nedevschi, 2014)	42.30%	DF	✓							Many	C
MT-DPM (Yan *et al.*, 2013)	40.54%	DPM					✓	✓		HOG	C
JointDeep (Ouyang and Wang, 2013a)	39.32%	DN			✓					Color+Gradient	C
SDN (Luo *et al.*, 2014)	37.87%	DN				✓	✓			Pixels	C
MT-DPM+Context (Yan *et al.*, 2013)	37.64%	DPM			✓		✓	✓		HOG	C+
ACF+SDt (Park *et al.*, 2013)	37.34%	DF	✓						✓	ACF+Flow	C+
SquaresChnFtrs (Benenson *et al.*, 2013)	34.81%	DF	✓							HOG+LUV	C
InformedHaar (Zhang *et al.*, 2014)	34.60%	DF	✓							HOG+LUV	C
Katamari-v1 (ours)	22.49%	DF	✓		✓				✓	HOG+Flow	C+

Table 4.1: List of methods with *Caltech* results sorted by log-average miss rate (lower is better). Consult Sec. 4.3 for descriptions of each column. See also matching Fig. 4.3. "*HOG*" refers to our re-implementation of (Dalal and Triggs, 2005).

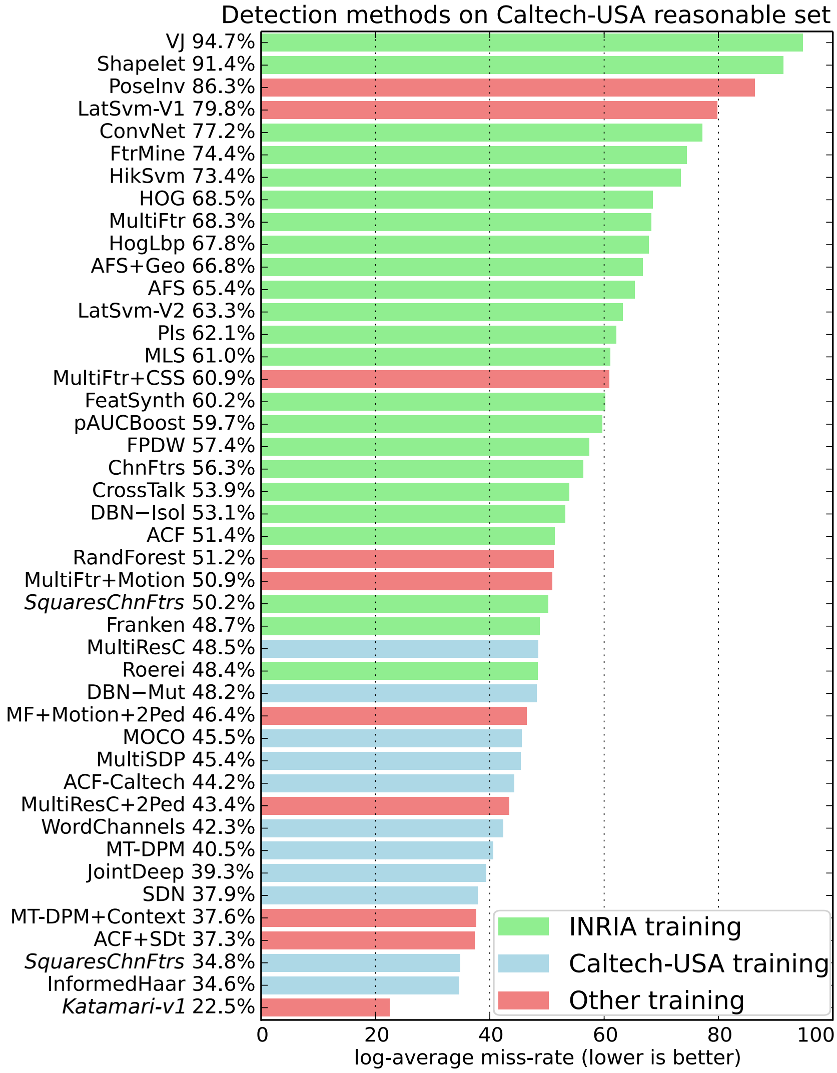


Figure 4.3: *Caltech* detection results.

additional details on the training data used[2]. High performing methods with "other training" use extended versions of *Caltech*. For instance *MultiResC+2Ped* uses *Caltech* plus an extended set of annotations over *INRIA*, *MT-DPM+Context* uses an external training set for cars, and *ACF+SDt* employs additional frames from the original *Caltech* videos.

4.3.2 Solution Families

Overall we notice that out of the 40+ methods we can discern three families: 1) *DPM* variants (e.g. *MultiResC* (Park *et al.*, 2010), *MT-DPM* (Yan *et al.*, 2013)), 2) Deep

[2] "Training" data column: I→INRIA, C→Caltech, I+/C+ →INRIA/Caltech and additional data, P→Pascal, T→TUD-Motion, I&C→both *INRIA* and *Caltech*.

networks (e.g. *JointDeep* (Ouyang and Wang, 2013a), *ConvNet* (Sermanet *et al.*, 2013)), and 3) Decision forests (e.g. *ChnFtrs*, *Roerei* (Benenson *et al.*, 2013)). In Tab. 4.1 we identify these families as DPM, DN, and DF respectively.

Based on raw numbers alone, boosted decision trees (DF) seem particularly suited for pedestrian detection, reaching top performance on both the "train on *INRIA*, test on *Caltech*", and "train on *Caltech*, test on *Caltech*" tasks. It is unclear, however, what gives them an edge. The deep networks explored also show interesting properties and fast progress in detection quality.

Conclusion. Overall, among the methods compared, *DPM* variants, deep networks, and (boosted) decision forests all reach top performance in pedestrian detection (around 37 % laMR on *Caltech*, see Fig. 4.3).

4.3.3 Better Classifiers

Since the original proposal of *HOG+SVM* (Dalal and Triggs, 2005), linear and non-linear kernels have been considered. *HikSvm* (Maji *et al.*, 2008) considered fast approximations of non-linear kernels. This method obtains improvements when using the flawed FPPW evaluation metric (see Sec. 4.3), but fails to perform well under the proper evaluation (FPPI). In the work on *MultiFtrs* (Wojek and Schiele, 2008), it was argued that given enough features, *AdaBoost* and linear SVMs perform at roughly similar levels for pedestrian detection.

Recently, more and more components of the detector are optimized jointly with the "decision component" (e.g. pooling regions in *ChnFtrs* (Dollár *et al.*, 2009a), filters in *JointDeep* (Ouyang and Wang, 2013a)). As a result the distinction between features and classifiers is not clear-cut any more (see also Sections 4.3.8 and 4.3.9). This is a trend that has continued with modern day end-to-end-trained detectors.

Conclusion. There is no conclusive empirical evidence indicating whether non-linear kernels provide meaningful gains over linear kernels when using non-trivial features for pedestrian detection. Similarly, it is unclear whether one particular type of classifier (e.g. SVM or decision forests) is better suited for pedestrian detection than another. Recent results appear to bear this out, as modern detectors rely on simple linear classifiers together with powerful learned features.

4.3.4 Additional Data

The core problem of pedestrian detection focuses on individual monocular RGB images. Some methods explore the use of additional information at training and test time to improve detections. They consider stereo images (Keller *et al.*, 2011), optical flow (using

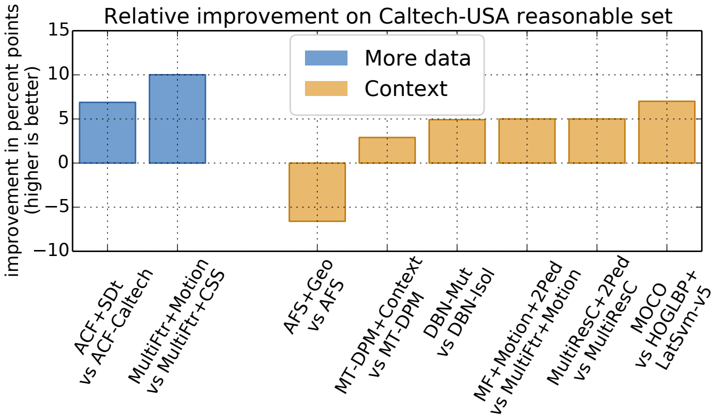


Figure 4.4: *Caltech* detection improvements for different method types. The improvement is reported relative to each method's relevant baseline as indicated by the labels of the x-axis ("method vs. baseline").

preceding frames, e.g. *MultiFtr+Motion* (Walk *et al.*, 2010) and *ACF+SDt* (Park *et al.*, 2013)), tracking (Ess *et al.*, 2009), or data from other sensors (such as lidar (Premebida *et al.*, 2014) or radar).

For monocular methods it is still unclear how much tracking can improve per-frame detection itself. As seen in Fig. 4.4 exploiting optical flow provides a non-trivial improvement over the baselines. Curiously, the top results at the time of publication of this work (*ACF-SDt*, Park *et al.*, 2013) are obtained using coarse rather than high quality flow. In Sec. 4.4.2 we examine the complementarity of flow with other ingredients. Good success exploiting flow and stereo on the *Daimler* dataset has been reported (Enzweiler and Gavrila, 2011), but similar results have yet to be attained on newer datasets such as *KITTI*.

Conclusion. At the time of the initial publication of this study (Benenson *et al.*, 2014), using additional data provides meaningful improvements. However, on modern datasets stereo and flow cues have yet to be fully exploited. Methods that merely rely on single monocular frames have been able to keep up with the performance improvements introduced by the use of auxiliary information whether depth or motion.

4.3.5 Exploiting Context

Sliding window detectors score potential detection windows using the content inside that window. Drawing on the context of the detection window, i.e. image content surrounding the window, can improve detection performance. Strategies for exploiting context include: ground plane constraints (*MultiResC* (Park *et al.*, 2010), *RandForest* (Marín *et al.*, 2013)), variants of auto-context (Tu and Bai (2010), *MOCO* (Chen *et al.*, 2013)), other category detectors (*MT-DPM+Context* (Yan *et al.*, 2013)), and person-to-person patterns (*DBN-Mut* (Ouyang *et al.*, 2013), *+2Ped* (Ouyang and Wang, 2013b), and *JointDeep* (Ouyang and Wang, 2013a)).

Fig. 4.4 shows the performance improvement for methods incorporating context. Overall, we see improvements in absolute terms of 3% – 7% laMR. The negative impact of *AFS+Geo* is due to the use of the updated evaluation metric (FPPI vs. FPPW) — see Sec. 4.3. Interestingly, *+2Ped* (Ouyang and Wang, 2013b) obtains a consistent absolute improvement of 2% – 5% laMR over existing methods, even top performing ones (Sec. 4.4.2).

Conclusion. Context provides consistent improvements for pedestrian detection, although the amount of improvement is lower compared to additional test data (Sec. 4.3.4) and deep architectures (Sec. 4.3.8). The bulk of detection quality must come from other sources.

4.3.6 Part-based Models

DPM (Felzenszwalb *et al.*, 2010) was originally motivated for pedestrian detection. Modelling pedestrians with hierarchical part-based models that allow for flexible part compositions is an idea that has become very popular and dozens of variants have been explored. For pedestrian detection the results are competitive, but not noticeably stronger than other detector families. Variants include here are *LatSvm* (Yan *et al.*, 2014; Felzenszwalb *et al.*, 2008), *MultiResC* (Park *et al.*, 2010), and *MT-DPM* (Yan *et al.*, 2013). More interesting results have been obtained when modelling parts and their deformations inside a deep architecture, e.g. *DBN-Mut* (Ouyang *et al.*, 2013) and *JointDeep* (Ouyang and Wang, 2013a).

DPM and its variants are systematically outmatched by methods using a single component and no parts, e.g. *Roerei* (Benenson *et al.*, 2013) and *SquaresChnFtrs* (Sec. 4.4.1, casting doubt on the need for explicit part modelling. Recent work has explored ways to capture deformations entirely without part-specific detectors (Hariharan *et al.*, 2014b; Pedersoli *et al.*, 2014). Some work has even suggested that *DPM*'s use of multiple templates (or "components") per class is more critical to its success than the modelling of part deformations (Divvala *et al.*, 2012).

Conclusion. For pedestrian detection there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling. Rigid detectors perform just as well, but this is most likely a result of the pose distribution of pedestrians which can be captured adequately by rigid templates.

4.3.7 Multi-scale Models

Typically for detection, both high and low resolution candidate windows are resampled to a common size before extracting features. It has recently been noticed that training different models for different resolutions systematically improve absolute performance by 1% – 2% laMR (Park *et al.*, 2010; Benenson *et al.*, 2013; Yan *et al.*, 2013), since the detector has access to the full information available at each window size. This technique does not impact computational cost at detection time (Benenson *et al.*, 2012), although training time increases.

Conclusion. Multi-scale models provide a simple and generic extension to existing detectors. Despite consistent improvements, their contribution to the final quality is rather minor overall.

4.3.8 Deep Architectures

Large amounts of training data and increased computing power have lead to recent successes of deep architectures — typically convolutional neural networks (CNNs) — on diverse computer vision tasks, such as large-scale classification and detection (Krizhevsky *et al.*, 2012; Girshick *et al.*, 2014; Sermanet *et al.*, 2014), and semantic labelling (Pinheiro and Collobert, 2014). These results have inspired the application of deep architectures to the pedestrian task.

ConvNet (Sermanet *et al.*, 2013) uses a mix of unsupervised and supervised training to create a CNN trained on *INRIA*. This method obtains fair results on *INRIA*, *ETH*, and *TUD-Brussels*, however fails to generalise to *Caltech*. This method learns to extract features directly from raw pixel values.

Another line of work focuses on using deep architectures to jointly model parts and occlusions (*DBN-Isol* (Ouyang and Wang, 2012), *DBN-Mut* (Ouyang *et al.*, 2013), *JointDeep* (Ouyang and Wang, 2013a), and *SDN* (Luo *et al.*, 2014)). The absolute performance improvements of such models varies between 1.5% to 14% laMR. Note that these works use edge and colour features as inputs (Ouyang and Wang, 2013a; Ouyang *et al.*, 2013; Ouyang and Wang, 2012), or initialise network weights to edge-sensitive filters, rather than discovering features from raw pixel values as usually done in deep architectures. No results had yet been reported using features pre-trained on *ImageNet* at the time of publication, as in Girshick *et al.* (2014) and Azizpour *et al.* (2015). In Chapter 5, we present such results.

Conclusion. At the time of this study, deep networks had not yet extended their success at learning features to pedestrian detection. Neural networks were used to model higher-level aspects of people such as part relations occlusions, and context, while still relying on traditional feature extraction pipelines. The obtained results were on par with *DPM*- and decision-forest-based approaches. This has since changed drastically as we outline in Chapter 2. Neural networks have taken over this subdomain as well.

4.3.9 Better features

The most popular approach (about 30 % of the considered methods) for improving detection quality is to increase and diversify the features computed over the input image. By having richer and higher dimensional representations, the classification task becomes somewhat easier, enabling improved results. A large set of feature types have been explored: edge information (Dalal and Triggs, 2005; Dollár *et al.*, 2009a; Lim *et al.*, 2013; Luo *et al.*, 2014), colour information (Dollár *et al.*, 2009a; Walk *et al.*, 2010), texture information (Wang *et al.*, 2009), local shape information (Costea and Nedevschi, 2014), covariance features (Paisitkriangkrai *et al.*, 2013), among others. More and more diverse features have been shown to systematically improve performance.

While various decision forest methods use 10 feature channels (e.g. *ChnFtrs*, *ACF*, *Roerei*, *SquaresChnFtrs*), some papers have considered up to an order of magnitude more channels (Wojek and Schiele, 2008; Lim *et al.*, 2013; Paisitkriangkrai *et al.*, 2013; Marín *et al.*, 2013; Costea and Nedevschi, 2014). Despite the improvements obtained by adding more diverse channels, top performance can still reached with only 10 channels (6 gradient orientations, 1 gradient magnitude, and 3 colour channels, we name these *HOG+LUV* (see Tab. 4.1 and Fig. 4.3). In Sec. 4.4.1 we study different feature combinations in more detail.

From *VJ* (95% laMR) to *ChnFtrs* (56.3% laMR, by adding HOG and LUV channels), to *SquaresChnFtrs-Inria* (50.2% laMR, by exhaustive search over pooling sizes, see Sec. 4.4), improving feature representations drives progress. Switching training sets (Sec. 4.3.1) enables *SquaresChnFtrs-Caltech* to reach state of the art performance on *Caltech*, improving over significantly more sophisticated methods. *InformedHaar* (Zhang *et al.*, 2014) obtains top results by using a set of Haar wavelet-like features manually designed for the pedestrian detection task. In contrast *SquaresChnFtrs-Caltech* obtains similar results without using hand-crafted pooling regions.

More recent studies show that using more and better features yields further improvements (Paisitkriangkrai *et al.*, 2014; Nam *et al.*, 2014). It should be noted that better features for pedestrian detection had at this point not yet been obtained via deep learning approaches (see caveat on *ImageNet* features in Sec. 4.3.8).

Conclusion. In the decade preceding this analysis, improved features were a constant driver for detection quality improvement, suggesting that this would continue to be the case in the years that followed. Most of this improvement had been obtained by extensive trial and error in feature design. Our contention was that the next scientific step would be to develop a more profound understanding of the what makes good features good, and how to design or learn even better ones. Results that followed demonstrate that features continue to play an outsize role in driving performance gains in detection.

4.4 Experiments

Based on our analysis in the previous section, three aspects seem to be the most promising in terms of impact on detection quality: better features (Sec. 4.3.9), additional data (Sec. 4.3.4), and context information (Sec. 4.3.5). We thus conduct experiments on the complementarity of these aspects.

Among the three solution families discussed (Sec. 4.3.2), we choose the *Integral Channels Features* framework (Dollár *et al.*, 2009a) (a decision forest) for conducting our experiments. Methods from this family have shown good performance, train in minutes∼hours, and lend themselves to the analyses we aim.

In particular, we use the (open source) *SquaresChnFtrs* baseline described in (Benenson *et al.*, 2013): 2048 level-2 decision trees (3 threshold comparisons per tree) over *HOG+LUV* channels (10 channels), composing one 64 × 128 pixels template learned via vanilla *AdaBoost* and few bootstrapping rounds of hard negative mining.

4.4.1 How Much Do Features Matter?

In this section, we evaluate the impact of increasing feature complexity. We tune all methods on the *INRIA* test set, and demonstrate results on the *Caltech* test set (see Fig. 4.5).

The first series of experiments aims at mimicking landmark detection techniques, such as *VJ* (Viola *et al.*, 2003), *HOG+linSVM* (Dalal and Triggs, 2005), and *ChnFtrs* (Dollár *et al.*, 2009a). *VJLike* uses only the luminance colour channel, emulating the Haar wavelet-like features from the original (Viola *et al.*, 2003) using level-2 decision trees. *HOGLike-L1/L2* use 8 × 8 pixel pooling regions, 1 gradient magnitude and 6 oriented gradient channels, as well as level 1/2 decision trees. We also report results when adding the LUV colour channels *HOGLike+LUV* (10 feature channels total). *SquaresChnFtrs* is the baseline described in the beginning of Sec. 4.4, which is similar to *HOGLike+LUV* to but with square pooling regions of any size.

Inspired by Nam *et al.* (2014), we also expand the 10 HOG+LUV channels into 40 channels by convolving each channel with three DCT (discrete cosine transform) basis functions (of 7 × 7 pixels), and storing the absolute value of the filter responses as additional feature channels. We name this variant *SquaresChnFtrs+DCT*.

Conclusion. Much of the progress since *VJ* can by explained by the use of better features, based on oriented gradients and colour information. Simple tweaks to these

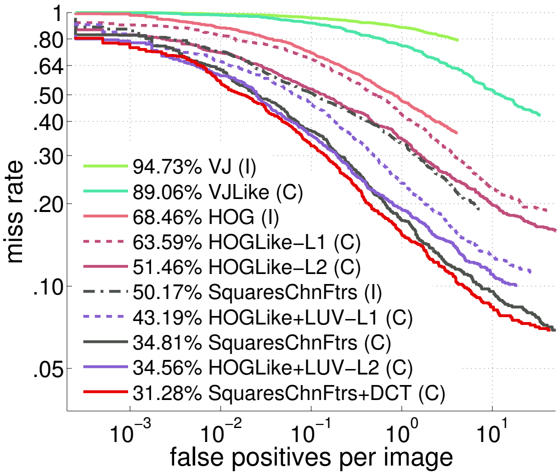


Figure 4.5: Effect of features on detection performance on the *Caltech* "Reasonable" test set.

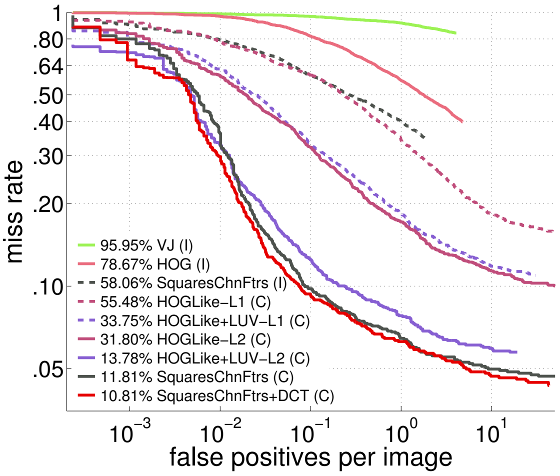


Figure 4.6: Caltech training set performance. (I)/(C) indicates the use of either *INRIA* or *Caltech* for training.

well known features (e.g. projection onto the DCT basis) can still yield noticeable improvements.

4.4.2 Complementarity of Detector Elements

After revisiting the effect of single frame features in Sec. 4.4.1 we now consider the complementarity of better features (HOG+LUV+DCT), additional data (via optical flow), and context (via person-to-person interactions).

We encode the optical flow using the same *SDt* features from *ACF+SDt* (Park *et al.*, 2010) (image difference between current frame T and coarsely aligned T-4 and T-8). The context information is injected using the *+2Ped* re-weighting strategy (Ouyang and Wang, 2013b) (the detection scores are combined with the scores of a "2 person" *DPM* detector). In all experiments both DCT and *SDt* features are pooled over 8×8 regions (as in Park *et al.* (2010)), instead of "all square sizes" for the HOG+LUV features.

We refer to our method combining *SquaresChnFtrs+DCT+SDt+2Ped* as *Katamari-v1*. It reaches the best performance on the *Caltech* dataset in 2014. In Fig. 4.7 we show it together with the best performing method for each training set and solution family at the time (see Tab. 4.1).

Conclusion. Our experiments show that extra single-frame features, motion features, and context information are largely complementary (12 % gain, instead of $3 + 7 + 5$ %), even when starting from a strong detector. It remains to be seen if future progress in detection quality will be obtained by further insights of the "core" algorithm (thus

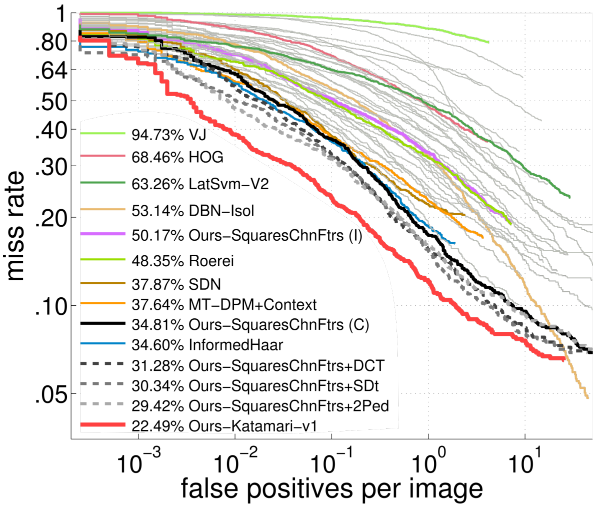


Figure 4.7: Some of the top quality detection methods for *Caltech*. See Sec. 4.4.2.

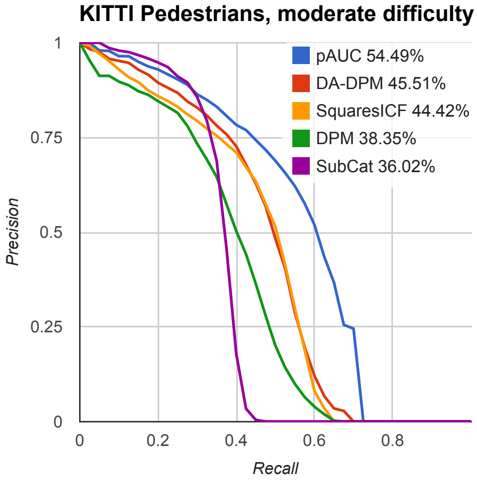


Figure 4.8: Pedestrian detection results on the *KITTI* dataset.

further diminishing the relative improvement of add-ons), or by extending the diversity of techniques employed inside a system.

4.4.3 How Much Model Capacity is Needed?

The main task of detection is to generalise from training to test set. Before we analyse the generalisation capability (Sec. 4.4.4), we consider a necessary condition for high quality detection: is the learned model performing well on the training set?

In Fig. 4.6 we see the detection quality of the models considered in Sec. 4.4.1, when evaluated over their training set. None of these methods performs perfectly on the training set. In fact, the trend is very similar to performance on the test set (see Fig. 4.5) and we do not observe yet symptoms of overfitting.

Conclusion. Our results indicate that research on increasing the discriminative power of detectors is likely to further improve detection quality. More discriminative power can originate from more and better features or more complex classifiers.

4.4.4 Generalisation across Datasets

For real world application beyond a specific benchmark, the generalisation capability of a model is key. In that sense results of models trained on *INRIA* and tested on *Caltech* are more relevant than the ones trained (and tested) on *Caltech*.

Training set Test set	INRIA	Caltech	KITTI
INRIA	*17.42* %	60.50 %	**55.83** %
Caltech	**50.17** %	*34.81* %	61.19 %
KITTI	**38.61** %	28.65 %	*44.42* %
ETH	**56.27** %	76.11%	61.19 %

Table 4.2: Effect of training set on the detection quality on different test sets. Bold indicates second best training set for each test set, except for *ETH* where bold indicates the best training set.

Tab. 4.2 shows the performance of *SquaresChnFtrs* over *Caltech* when using different training sets (laMR for *INRIA*/*Caltech*/*ETH*, AUC for *KITTI*. These experiments indicate that training on *Caltech* or *KITTI* provides little generalisation capability towards *INRIA*, while the converse is not true. Surprisingly, despite the visual similarity between *KITTI* and *Caltech*, *INRIA* is the second best training set choice for *KITTI* and *Caltech*. This shows that *Caltech* pedestrians are of "their own kind", and that training with *INRIA* is effective due to its diversity. In other words, a training set containing few diverse pedestrians (*INRIA*) is better than many similar ones (*Caltech*/*KITTI*).

The good news is that the best methods considered here seem to perform well both across datasets and when trained on the respective training data. Fig. 4.8 shows methods trained and tested on *KITTI*, and we see that *SquaresChnFtrs* (referred to here as *SquaresICF*) is better than vanilla *DPM* and on par with the best *DPM* variant. The best method on *KITTI* as of July 2014, *pAUC* (Paisitkriangkrai *et al.*, 2014), is a variant of *ChnFtrs* using 250 feature channels (see the *KITTI* website for details). These two observations are consistent with our discussions in Sections 4.3.9 and 4.4.1.

Conclusion. While detectors learned on one dataset may not necessarily transfer well to others, their ranking is stable across datasets, suggesting that insights can be learned from well-performing methods regardless of the benchmark.

4.5 Conclusions

Our experiments show that most of the progress in the preceding decade of pedestrian detection can be attributed to the improvement in features alone. Evidence suggests that this trend will continue. Although some of these features might be driven by learning, they are mainly hand-crafted via trial and error.

Our experiment combining the detector ingredients that our retrospective analysis found to work well (better features, optical flow, and context) shows that these ingredients are mostly complementary. Their combination produces best published detection performance on *Caltech* in July 2014.

While the three big families of pedestrian detectors (deformable part models, decision forests, deep networks) are based on different learning techniques, their state-of-the-art results are surprisingly close.

We also showed that cross-dataset generalisation is an issue, which requires research into better features and better domain adaptation.

The main challenge ahead seems to develop a deeper understanding of what makes good features good, so as to enable the design of even better ones.

5 Deep Learning for Pedestrian Detection

In Chapter 4 we saw that the progress of a decade of research on pedestrian detection has been driven by improvements in feature engineering. While representation learning for pedestrian detection had been explored by then, at that point the strongest detectors were still built on top of hand-crafted "feature channels".

In this chapter we study the use of convolutional neural networks (CNNs) that operate on raw pixel values and without the use of hand-crafted features. Despite their recent diverse successes, CNNs had not yet caught up to classical pedestrian detectors. Unlike competing work, we deliberately avoided explicitly modelling the problem into the network (e.g. by considering parts or occlusion handling) and show that this is adequate for competitive performance. In a wide range of experiments we analyse differently sized CNNs, various architectural choices, hyperparameters, and the influence of different training sets — including pre-training on surrogate tasks.

This work was published at CVPR (Hosang *et al.*, 2015) and Jan Hosang was the lead author. Mohamed Omran contributed the experiment that motivated this work — demonstrating that with the same training and test data a simple image classification network outperforms state-of-the-art domain-specific deep detectors — as well as the experiments on smaller networks. We present the best CNN detector on the *Caltech* and *KITTI* dataset at the time, improving over all previous CNNs both for the *Caltech*1× and *Caltech*10× training setup (see Sec. 2.2). Using additional data at training time, our strongest CNN model is competitive even with previous detectors that use additional data (optical flow) at test time.

5.1 Introduction

In recent years the field of computer vision has seen an explosion of success stories involving CNNs. Such architectures currently provide top results for general object classification (Krizhevsky *et al.*, 2012; Russakovsky *et al.*, 2015a; Szegedy *et al.*, 2015), general object detection (Girshick *et al.*, 2014), feature matching (Long *et al.*, 2014), stereo matching (Zbontar and LeCun, 2015), scene recognition (Zhou *et al.*, 2014; Chen *et al.*, 2014), pose estimation (Toshev and Szegedy, 2014; Tompson *et al.*, 2014), action recognition (Karpathy *et al.*, 2014; Simonyan and Zisserman, 2014) and many other tasks (Razavian *et al.*, 2014; Azizpour *et al.*, 2015). Here, our motivation is to apply to the task of pedestrian detection recent insights regarding the training of large CNNs.

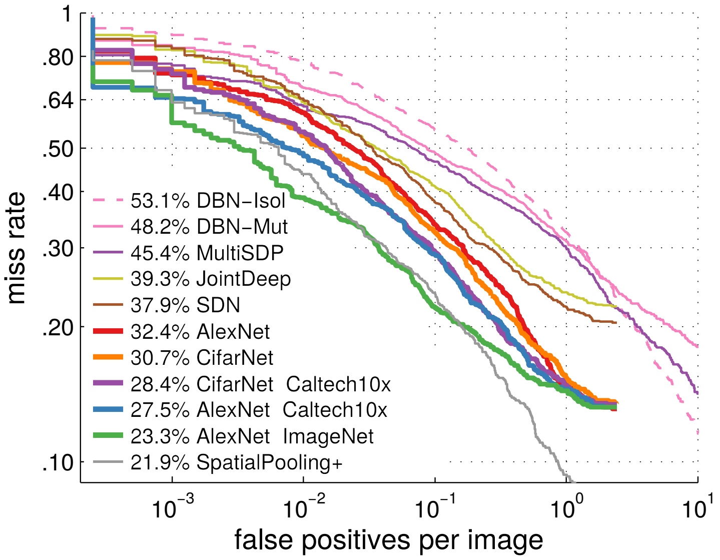


Figure 5.1: Comparison of CNN methods on the *Caltech* "Reasonable" test set (see Sec. 5.7). At the time of publication of this chapter, our *CifarNet* and *AlexNet* results significantly improved over previous CNNs, and matched the best reported results at that time (*SpatialPooling+*, which additionally uses optical flow).

Previous work on neural networks for pedestrian detection has relied on special-purpose designs, e.g. the use of hand-crafted features as inputs, part and occlusion modelling. Although these proposed methods perform reasonably, previous top methods are all based on decision trees learned via *AdaBoost* (e.g. Benenson *et al.*, 2014; Zhang *et al.*, 2014; Paisitkriangkrai *et al.*, 2014; Nam *et al.*, 2014; Wang *et al.*, 2013). This makes pedestrian detection an outlier among the many tasks enumerated above in which CNNs have left traditional methods in the dust performance-wise.

In this work we revisit the question, and show that both small and large vanilla CNNs can reach top performance on the challenging *Caltech* dataset (Dollár *et al.*, 2012b). We provide extensive experiments that cover training settings, network parameters, and different methods for generating object hypotheses or proposals.

Object detection. CNNs have been successfully applied to the task of generic object detection, showing strong results on datasets like *ImageNet* (Russakovsky *et al.*, 2015a; Krizhevsky *et al.*, 2012; He *et al.*, 2014; Szegedy *et al.*, 2015; Ouyang *et al.*, 2015; Simonyan and Zisserman, 2015) and *PASCAL VOC* (Girshick *et al.*, 2014; Agrawal *et al.*, 2014). The most successful generic object detectors are variants of the *R-CNN* framework (Girshick *et al.*, 2014). Given an input image, a sparse set of hypotheses are generated by a separate method. These so-called object proposals are subsequently classified via a CNN. This is essentially a two-stage cascade sliding window method.

The most popular proposal method for generic objects at the time of publication was *SelectiveSearch* (Uijlings *et al.*, 2013), which was also used with the original *R-CNN* detector (Girshick *et al.*, 2014). A fast and effective alternative is *EdgeBoxes* (Zitnick and Dollár, 2014). However, pedestrian detection methods based on neural networks (NNs) use classical detectors as a first stage proposal generator. *DBN-Isol* and *DBN-Mut* for example use *DPM* (Felzenszwalb *et al.*, 2010), while *JointDeep*, *MultiSDP*, and *SDN* rely on a *HOG+CSS+linearSVM* detector similar to the method of Walk *et al.* (2010). Only *ConvNet* (Sermanet *et al.*, 2013) applies a CNN in a sliding window fashion to the raw input image. For a more detailed comparison of these early NN-based pedestrian detector methods see Sec. 2.3.3, and for a more complete discussion of proposal methods, we refer the interested reader to the survey of Hosang *et al.* (2016).

Decision forests. Until 2015, most proposed methods for pedestrian detection did not use CNNs, relying instead on hand-crafted features. Focusing on single-frame methods, the top performing methods (on Caltech and *KITTI*) at the time of publication were *SquaresChnFtrs* (Chapter 4), *InformedHaar* (Zhang *et al.*, 2014), *SpatialPooling+* (Paisitkriangkrai *et al.*, 2014), *LDCF* (Nam *et al.*, 2014), and *Regionlets* (Wang *et al.*, 2013). All of them consist of boosted decision forests and can be considered variants of the integral channels features architecture (Dollár *et al.*, 2009a). *Regionlets* and *SpatialPooling+* use a large set of features, including HOG, LBP and CSS, while *SquaresChnFtrs*, *InformedHaar*, and *LDCF* build on HOG+LUV. On the *Caltech* benchmark, the previously best CNN, *SDN*, had been outperformed by all aforementioned methods.[3]

Input to CNNs. It is also important to highlight that *ConvNet* (Sermanet *et al.*, 2013) learns to predict from YUV input pixels, whereas all other methods use additional hand-crafted features. *DBN-Isol* and *DBN-Mut* use HOG features as input. *MultiSDP* uses HOG+CSS features as input. *JointDeep* and *SDN* use YUV+Gradients as input (and HOG+CSS for the detection proposals). We will show in our experiments that good performance can be reached using RGB inputs alone, but we also show that more sophisticated inputs systematically improve detection quality. Our data indicates that end-to-end features still have room for improvement and do not fully make hand-crafted features obsolete.

5.1.1 Contributions

In this chapter we propose to revisit pedestrian detection with CNNs by carefully exploring the design space (e.g. number of layers, filter sizes), and critical implementation choices (e.g. training data preprocessing, effect of detection proposals). We show that both small (10^5 parameters) and large ($6 \cdot 10^7$ parameters) networks can reach good performance when trained from scratch (even when using the exact same training data

[3] *Regionlets* matches *SpatialPooling+* on the *KITTI* benchmark, and based on our results in the previous chapter showing that rankings of methods are mostly preserved across datasets, would most likely improve over *SDN* on *Caltech* as well.

as previous methods). We also show the benefits of using extended and external data, which leads to the strongest single-frame detector on *Caltech* at the time of this study. At the time of publication, we report the best known performance for a CNN on the challenging *Caltech* dataset (improving by more than 10% in absolute terms) and the first CNN results on the *KITTI* dataset.

5.2 Training Data

It is well known that modern deep CNNs can exploit large amounts the training data and that this is in fact critical for performance. We will make use of two types of datasets: pedestrian detection datasets as well as large-scale image classification datasets for the purpose of pre-training our CNNs.

Pedestrian Detection: Caltech & KITTI. In Sec. 2.2, we describe both datasets in detail, but here we would like to describe the dataset splits we use for validation. For the *Caltech* experiments we use one of the validation splits suggested by Dollár *et al.* (2012b): the first five training videos are used for validation training and the sixth training video for validation testing. With *KITTI* (Geiger *et al.*, 2012), we split the public training set into train/validation (~4k/2k images) sets.

Large-scale Classification: *ImageNet* & Places. In Sec. 5.5 we will consider using large CNNs that can benefit from pre-training for surrogate tasks. We consider two image classification datasets: the *ImageNet ILSVRC2012* classification benchmark (object classification with 1000 categories) (Krizhevsky *et al.*, 2012; Russakovsky *et al.*, 2015a) and the *Places* dataset (scene classification with 205 categories) (Zhou *et al.*, 2014). The datasets provide $1.2 \cdot 10^6$ and $2.5 \cdot 10^6$ annotated images for training respectively.

5.3 From Decision Forests to Neural Networks

Before describing our experimental results, it is worth noting that the proposal method we are using — *SquaresChnFtrs* (see Sec. 5.4.1) — can be converted into a CNN. The overall system then becomes a cascade of two neural networks. *SquaresChnFtrs* (Chapter 4, Benenson *et al.* (2013)) is a decision forest composed of 2 048 level-2 decision trees, applied to ten hand-crafted feature channels (HOG+LUV). Rectangular regions of these feature channels are sum-pooled and fed to the split nodes of the trees. This architecture can be mapped to a CNN. Older work exploring this connection includes Sethi and Otten (1990); Cios and Ning (1992); Ivanova and Kubat (1995); Banerjee (1994); Setiono and Leow (1999).

As mentioned in Sec. 4.3.8, using hand-crafted features as inputs was common practice among early CNN-based pedestrian detectors (more on this in Sec. 5.4.4). We thus examined the possibility of initialising a neural network with the parameters of hand-crafted detector, since the operations involved can be straightforwardly mapped to standard neural network building blocks. The sum-pooling stage maps directly to an inner product layer. Each decision tree maps to a small column of two hidden layers, with sign-function non-linearities (hard non-linearities). Finally the output of all trees is combined via linear weighting.

The mapping from *SquaresChnFtrs* to a deep neural network is exact: evaluating the same inputs results in the exact same outputs. What is special about the resulting network is that it has not been trained by backpropagation but via *AdaBoost*. This network already performs better than the previously best CNN on *Caltech*, *SDN* (Luo *et al.*, 2014). Unfortunately however, experiments to soften the non-linearities and use backpropagation to fine-tune the model parameters did not show significant improvements. We suspect that the parameters found via *AdaBoost* are a local minimum that is hard to escape via stochastic gradient descent.

5.4 Vanilla Convolutional Neural Networks

In our experience many CNN architectures and training hyperparameters do not enable effective learning for diverse and challenging tasks. Following best practices, we thus start our exploration from architectures and parameters that are known to work well and progressively adapt them to the task at hand. In this section we thus first consider *CifarNet*, a small network designed to solve the *CIFAR-10* classification problem (10 objects categories, $(5+1)\cdot 10^5$ colour images of 32×32 pixels) (Krizhevsky, 2009). In Sec. 5.5 we consider *AlexNet*, a network that has 600 times more parameters than *CifarNet* and designed to solve the *ILSVRC2012* classification problem (1 000 objects categories, $(1.2+0.15)\cdot 10^6$ colour images of ∼VGA resolution). Both of these networks were introduced in Krizhevsky *et al.* (2012) and are re-implemented in as part of the open source *Caffe* project (Jia *et al.*, 2014)[4].

Although pedestrian detection is quite a different task than *CIFAR-10*, we decide to start our exploration from *CifarNet*, which provides fair performance on *CIFAR-10*. Its architecture is depicted in Fig. 5.2, and unless otherwise specified we use raw RGB inputs. We first discuss how to use the *CifarNet* network (Sec. 5.4.1). This naive approach already improves over the previously best CNNs for pedestrian detection (Sec. 5.4.2). Sections 5.4.3 and 5.4.4 explore the design space around *CifarNet* and further push the detection quality. All models in this section are trained using *Caltech* data only (see Sec. 5.2).

[4] http://caffe.berkeleyvision.org

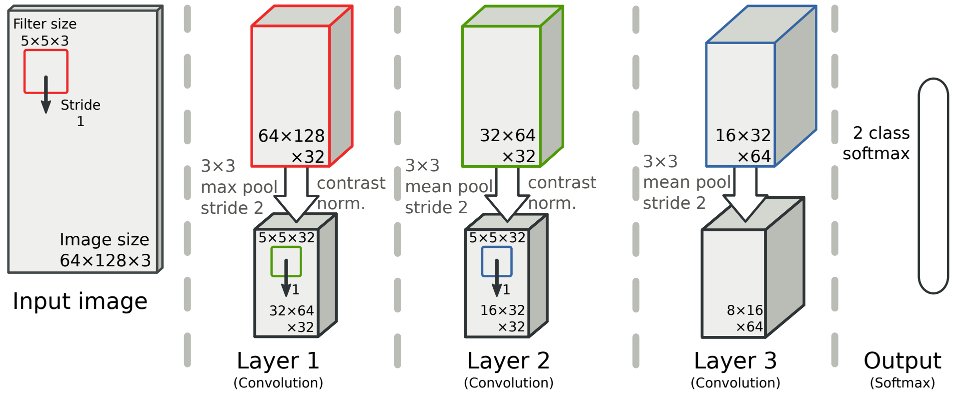


Figure 5.2: Illustration of the *CifarNet* architecture, $\sim 10^5$ parameters.

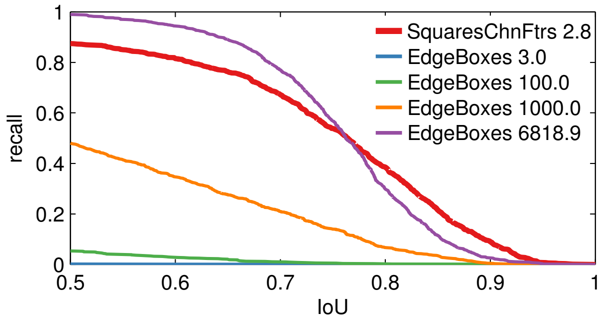


Figure 5.3: Recall of ground truth annotations versus Intersection-over-Union threshold on the *Caltech* test set. The legend indicates the average number of detection proposals per image for each curve. A pedestrian detector generates much better proposals than a state of the art generic method (*EdgeBoxes* (Zitnick and Dollár, 2014)).

5.4.1 How to use CifarNet?

Given an initial network specification, there are still several design choices that affect the final detection quality. We discuss some of them in the following paragraphs.

Detection proposals. Unless otherwise specified we use the *SquaresChnFtrs* (Chapter 4) detector to generate proposals because, at the time of publication, it was the best performing pedestrian detector on *Caltech* with accessible source code. In Fig. 5.3 we compare *SquaresChnFtrs* against *EdgeBoxes* (Zitnick and Dollár, 2014), a state of the art class-agnostic proposal method. Using class-specific proposals allows us to reduce the number of proposals by three orders of magnitude. Other than *ConvNet* (Sermanet *et al.*, 2013) which does not use proposals, all other competing CNNs also use a pedestrian detector for proposals (see also Sec. 5.4.2).

Positives	Negatives	laMR
GT	Random	83.1%
GT	`IoU` < 0.5	37.1%
GT	`IoU` < 0.3	37.2%
GT, `IoU` > 0.5	`IoU` < 0.5	42.1%
GT, `IoU` > 0.5	`IoU` < 0.3	41.3%
GT, `IoU` > 0.75	`IoU` < 0.5	39.9%

Table 5.1: Effect of positive and negative training sets on the detection quality. laMR: log-average miss rate on *Caltech* validation set. GT: ground truth bounding boxes.

Window size	laMR
32 × 32	50.6%
64 × 32	48.2%
128 × 64	39.9%
128 × 128	49.4%
227 × 227	54.9%

Table 5.2: Effect of window size on performance. (laMR: see Tab. 5.1)

Ratio	laMR
None	41.4%
1 : 10	40.6%
1 : 5	39.9%
1 : 1	39.8%

Table 5.3: Performance as function of strictly enforced ratio of positives:negatives in each training batch. *None*: none enforced. (laMR: see Tab. 5.1)

Thresholds for positive and negative samples. Given both training proposals and ground truth (GT) annotations, we now consider which training label to assign to each proposal. A proposal is considered to be a positive example if it exceeds a certain Intersection over Union (IoU) threshold for at least one GT annotation. It is considered negative if it does not exceed a second IoU threshold for any GT annotation, and is ignored otherwise. We find that using GT annotations as positives is beneficial (i.e. not applying significant jitter, see Tab. 5.1).

Model window size. A typical choice for pedestrian detectors is a model window size of 128×64 pixels in which the pedestrian occupies an area of 96×48 (Dalal and Triggs, 2005; Dollár *et al.*, 2009a; Benenson *et al.*, 2013, 2014). It is unclear that this is the ideal input size for CNNs. Despite *CifarNet* being designed to operate over 32×32 pixels, Tab. 5.2 shows that a model size of 128×64 pixels indeed works best. We experimented with other variants that involved stretching the proposals to different aspect or adding more context, but these led to no clear improvement.

Training batch. In a detection setup, training samples are typically highly imbalanced towards the background class. Although in our validation setup the imbalance is limited, we found it beneficial throughout our experiments to enforce a strict ratio of positive to negative examples per batch of the stochastic gradient descent optimisation (see Tab. 5.3). The final performance is not sensitive to this parameter as long as some fixed ratio (vs. *None*) is maintained. We use a ratio of 1 : 5.

Method	Proposals	Test laMR
Proposals of JointDeep	-	45.5%
JointDeep	Proposals of JointDeep	39.3%
SDN		37.9%
CifarNet		36.5%
SquaresChnFtrs	-	34.8% (Chapter 4)
CifarNet	SquaresChnFtrs	*30.7%*

Table 5.4: Detection quality as a function of the method and the proposals used for training and testing (laMR: log-average miss rate on *Caltech* test set). When using the exact same training data as *JointDeep* (Ouyang and Wang, 2013a), our vanilla *CifarNet* already improves over the previous best known CNN on *Caltech* (*SDN*, Luo *et al.* 2014).

5.4.2 How far can we get with CifarNet?

Given the parameter selection on the validation set from previous sections, how does *CifarNet* compare to previous CNN results on the *Caltech* test set? Tab. 5.4 and Fig. 5.1 show that our naive network immediately improves over the previously best CNN: 30.7% vs. 37.9% laMR (*SDN*, Luo *et al.* 2014).

To decouple the contribution of our strong *SquaresChnFtrs* proposals from the classification performance of *CifarNet*, we also train *CifarNet* with the exact same proposals used by *JointDeep* (Ouyang and Wang, 2013a). When using these both at training and test time (provided together with the official implementation), the vanilla *CifarNet* already improves over both the custom-designed *JointDeep* and *SDN*. Our *CifarNet* results are surprisingly close to the previously best known pedestrian detector trained on *Caltech*1×: 30.7% vs. 29.2% laMR (*SpatialPooling*, Paisitkriangkrai *et al.* 2014).

5.4.3 Exploring different architectures

Encouraged by our initial results, we proceed to explore different parameters for the *CifarNet* architecture.

Number and size of convolutional filters. Using the *Caltech* validation set we perform a sweep over different convolutional filter sizes (3×3, 5×5, or 7×7 pixels) and number of filters at each layer (16, 32, or 64 filters). We observe that using large filters hurts quality, while the varying the number of filters shows less impact. Although some fluctuation in log-average miss rate is observed, overall there is no clear trend indicating that a configuration is clearly better than another. For the sake of simplicity, we thus keep using *CifarNet* (32-32-64 filters of 5×5 pixels) in subsequent experiments.

# layers	Architecture	laMR
3	CONV1 CONV2 FC	47.6%
	CONV1 CONV2 LC	43.2%
	CONV1 CONV2 CONV3 (*CifarNet*, Fig. 5.2)	*37.1%*
4	CONV1 CONV2 CONV3 FC	39.6%
	CONV1 CONV2 CONV3 LC	40.5%
	CONV1 CONV2 FC1 FC2	43.2%
	CONV1 CONV2 CONV3 CONV4	43.3%
4	CONV1 CONV2 CONV3 CONCAT23 FC	38.4%

Table 5.5: Performance of different architectures, sorted by the number of layers before the softmax classifier. CONCAT23: concatenates CONV2 and CONV3 and passes on the resulting feature map. laMR: log-average miss rate on *Caltech* validation set

Input channels	# channels	CifarNet
RGB	3	39.9%
LUV	3	46.5%
G+LUV	4	40.0%
HOG+L	7	36.8%
HOG+LUV	10	40.7%

Table 5.6: Detection quality for different input configurations. G indicates luminance channel gradient, HOG indicates G plus G spread over six orientation bins (hard-binning). These are the same input channels used by our *SquaresChnFtrs* proposal method. Results in laMR (log-average miss rate) on *Caltech* validation set.

Number and type of layers. In Tab. 5.5 we evaluate the effect of changing the number and type of layers, while keeping other *CifarNet* parameters fixed. Besides convolutional layers (CONV) and fully-connected layers (FC), we also consider locally-connected layers (LC) (Taigman *et al.*, 2014), and concatenating features across layers (CONCAT23) (used in *ConvNet*, Sermanet *et al.* 2013). None of the considered architectural changes improve over the original *CifarNet*.

5.4.4 Input Channels

As mentioned above, the majority of previous CNNs for pedestrian detection use gradient and colour features as inputs, instead of raw RGB values. In Tab. 5.6 we evaluate the effect of different input features over *CifarNet*. It seems that HOG+L channels provide a small advantage over RGB. To allow for direct comparisons with the large networks we consider later, in the next sections we continue to use raw RGB as the input for our *CifarNet* experiments. We report *CifarNet* test set results in Sec. 5.6.

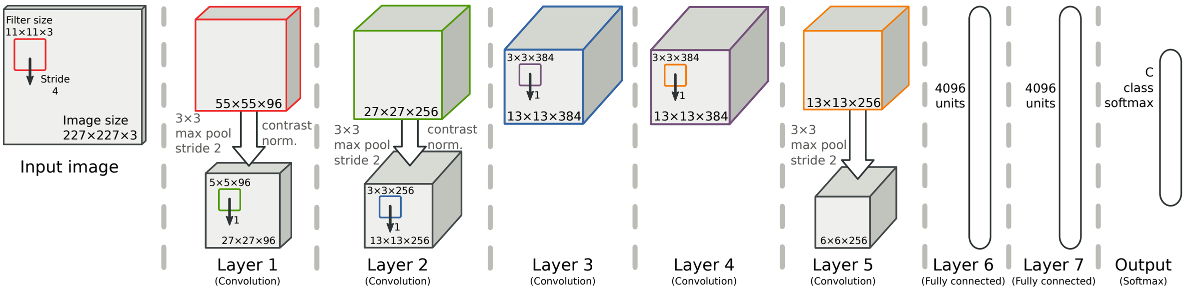


Figure 5.4: Illustration of the *AlexNet* architecture, $\sim 6 \cdot 10^7$ parameters.

5.5 Large Convolutional Neural Networks

An appealing characteristic of CNNs is their ability to leverage large amounts of training data. Batch training with stochastic gradient descent also avoids the prohibitive memory requirements of other classifiers. We now explore larger CNNs trained with more data.

We base our experiments on the *R-CNN* detector (Girshick *et al.*, 2014), which at the original time of writing was a top-performing method on the *PASCAL VOC* detection benchmark (Everingham *et al.*, 2015). We are thus interested in evaluating its performance on pedestrian detection.

5.5.1 Surrogate tasks for improved detections

The *R-CNN* approach ("Regions with CNN features") relies on *AlexNet* (see Fig. 5.4), a large network pre-trained for the *ImageNet* classification task (Krizhevsky *et al.*, 2012). We use "*AlexNet*" as shorthand for "*R-CNN* with *AlexNet*" with the distinction made clear by the context. During *R-CNN* training *AlexNet* is fine-tuned for the detection task, and in a second step, the softmax classifier is replaced by a linear SVM. Unless otherwise specified, we use the default parameters of the open source, *Caffe*-based *R-CNN* implementation[5]. Like in the previous sections, we use *SquaresChnFtrs* to produce detection proposals. For consistency with other *AlexNet* experiments in the literature, we use the default RGB and 227 × 227 input size, also noting that the optimal *CifarNet* parameters might not apply to the larger *AlexNet*.

Pre-training. If we only train the SVM classifier without fine-tuning the lower layers of *AlexNet*, we obtain 39.8% laMR on the *Caltech* test set. This is already surprisingly close to the result (37.9% laMR) of the previous best CNN for the task (*SDN*). When fine-tuning all layers on *Caltech*, the test set performance increases dramatically, reaching 25.9% laMR. This confirms the effectiveness of the general *R-CNN* recipe for detection (train *AlexNet* on *ImageNet*, then fine-tune for the task of interest).

[5] https://github.com/rbgirshick/rcnn

AlexNet training	Fine-tuning	SVM training	Test laMR
Random	none	Caltech1x	86.7%
ImageNet	none	Caltech1x	39.8%
Places+Imagenet	Caltech1x	Caltech1x	30.1%
Places			27.0%
ImageNet			25.9%
ImageNet	Positives10x	Positives10x	23.8%
	Caltech10x	Caltech10x	*23.3%*
Caltech1x	-	Caltech1x	32.4%
	-	Caltech10x	32.2%
Caltech10x	-	Caltech1x	*27.4%*
	-	Caltech10x	*27.5%*
SquaresChnFtrs (Chapter 4)			34.8%

Table 5.7: Detection quality when using different training data in different training stages of *AlexNet*: initial training of the CNN, optional fine-tuning of the CNN, and the SVM training. Positives10x: positives from *Caltech*10× and negatives from *Caltech*1×. Detection proposals provided by *SquaresChnFtrs*, result included for comparison. See Sections 5.5.1 and 5.5.2 for details.

In Tab. 5.7 we investigate the influence of the pre-training task by considering instances of *AlexNet* that have been trained for scene recognition (Zhou *et al.*, 2014) ("*Places*", see Sec. 5.2) and on both *Places* and *ImageNet*. "*Places*" provides results close to *ImageNet*, suggesting that the exact pre-training task is not critical and that there is nothing special about *ImageNet*.

Caltech10x. Due to the large number of parameters of *AlexNet*, we consider providing additional training data using *Caltech*10× for fine-tuning the network (see Sec. 5.2). Despite the strong correlation across training samples, we do observe further improvement (see Tab. 5.7). Interestingly, the bulk of the improvement is due to more pedestrians (*Positives*10× uses positives from *Caltech*10× and negatives from *Caltech*1×). Our top result, 23.3% laMR, makes our *AlexNet* setup the best reported single-frame detector on *Caltech* (i.e. without using optical flow) at the time of publication.

5.5.2 Caltech-only training

To compare with *CifarNet*, and to verify whether pre-training is necessary at all, we train *AlexNet* "from scratch" solely using the *Caltech* training data (Tab. 5.7).

Parameters	fc7	fc6	pool5	conv4
Default	32.2%	32.5%	33.4%	42.7%
Best	32.0%	31.8%	32.5%	42.4%

Table 5.8: Detection quality when training the *R-CNN* SVM over different layers of the finetuned CNN. "Best parameters" are found by exhaustive search on the validation set. laMR: log-average miss rate on *Caltech* validation set.

Training *AlexNet* solely on *Caltech* yields 32.4% laMR, which improves over the proposals (*SquaresChnFtrs*, 34.8% laMR) and the previous best known CNN on *Caltech* (*SDN*, 39.8% laMR). Using *Caltech*10× further pushes this down to 27.5% laMR.

Although these numbers are inferior to the ones obtained with an ImageNet pre-trained model (23.3% laMR, see Tab. 5.7), we can get surprisingly competitive results using only pedestrian data with randomly initialised network parameters despite the 10^7 free parameters of the *AlexNet* model. At the time we published this study (Hosang *et al.*, 2015), *AlexNet* with *Caltech*10× was the second best single-frame pedestrian detector to only use *Caltech* data — behind *LDCF* (24.8% laMR), which also uses *Caltech*10×).

5.5.3 Additional experiments

How many layers? So far all experiments use the default parameters of *R-CNN*. Previous works have reported that, depending on the task, using features from lower *AlexNet* layers can provide better results (e.g. Razavian *et al.*, 2014; Agrawal *et al.*, 2014; Azizpour *et al.*, 2015). Tab. 5.8 reports *Caltech* validation results when training the SVM output layer on top of layers four to seven (see Fig. 5.4). We report results when using the default parameter settings and parameters that have been optimised via grid search. These parameters are the SVM regularisation parameter as well as the criterion for choosing negative examples (upper bound on IoU with a ground truth example).

We observe a negligible difference between default and optimised parameters (at most -1%). Results for default parameters exhibit a slight trend of better performance for higher levels. These validation set results indicate that the *R-CNN* default parameters are a good choice overall for pedestrian detection.

Effect of proposal method. When comparing the performance of the proposal method compared to *AlexNet* fine-tuned on *Caltech*1×, we see an improvement of 9 pp (percentage points) in miss rate. In Tab. 5.9 we show the impact of using weaker or stronger proposals. Both *ACF* (Dollár *et al.*, 2014) and *SquaresChnFtrs* (Chapter 4, Benenson *et al.* 2013) provide source code, allowing us to generate training proposals. *Katamari* (Chapter 4) and *SpatialPooling+*(Paisitkriangkrai *et al.*, 2014) are top performers on the *Caltech* dataset, both using optical flow, i.e. additional information at test time. There is a

Fine-tuning	Training proposals	Testing proposals	Test laMR	Δ vs. proposals
1×	ACF	ACF	34.5%	9.7%
	SCF	ACF	34.3%	9.9%
	ACF	SCF	26.9%	7.9%
	SCF	SCF	25.9%	8.9%
	ACF	Katamari	25.1%	−2.6%
	SCF	Katamari	24.2%	−1.7%
10×	SCF	LDCF	23.4%	1.4%
	SCF	SCF	23.3%	11.5%
	SCF	SP+	22.0%	−0.1%
	SCF	Katamari	21.6%	0.9%
ACF (Dollár *et al.*, 2014)			44.2%	
SCF (SquaresChnFtrs, Chapter 4)			34.8%	
LDCF (Nam *et al.*, 2014)			24.8%	
Katamari (Chapter 4)			22.5%	
SP+ (SpatialPooling+, Paisitkriangkrai *et al.* 2014)			21.9%	

Table 5.9: Effect of proposal methods on detection quality of *R-CNN*. 1×/10× indicates fine-tuning on *Caltech*1× or *Caltech*10×. The last section of the table contains reference results for competing methods. Test laMR: log-average miss rate on *Caltech* test set. Δ: the improvement in laMR of the rescored proposals over the test proposals alone.

∼10 pp gap between the detectors *ACF*, *SquaresChnFtrs*, and *Katamari*/*SpatialPooling+*, allowing us to cover different operating points.

The results in Tab. 5.9 indicate that, despite the 10 pp gap, there is no noticeable difference between *AlexNet* models trained with *ACF* or *SquaresChnFtrs*. It is seems that as long as the proposals are not random (see top row of Tab. 5.1), the obtained quality is rather stable. The results also indicate that the quality improvement from *AlexNet* saturates around ∼22% laMR. Using stronger proposals does not lead to further improvement. This means that the discriminative power of our trained *AlexNet* is on par with the previously best known models on *Caltech*, but does not overtake them.

KITTI test set. In Fig. 5.5 we show *AlexNet* performance on the *KITTI* pedestrian detection benchmark (Geiger *et al.*, 2012). The network is pre-trained on *ImageNet* and fine-tuned using *KITTI* training data. *SquaresChnFtrs* reaches 44.4% AP (average precision), which *AlexNet* can improve to 50.1% AP. These are the earliest published results for CNNs on *KITTI*.

Given the ranking changes w.r.t. *Caltech* esp. of *SpatialPooling+*, it should be noted that (i) the two datasets use different evaluation metrics, (ii) the two datasets are more dissimilar than they seem on the surface (see Tab. 4.2), and (iii) overall *AlexNet* results on *KITTI* are satisfactory but proposals with higher recall might further improve results.

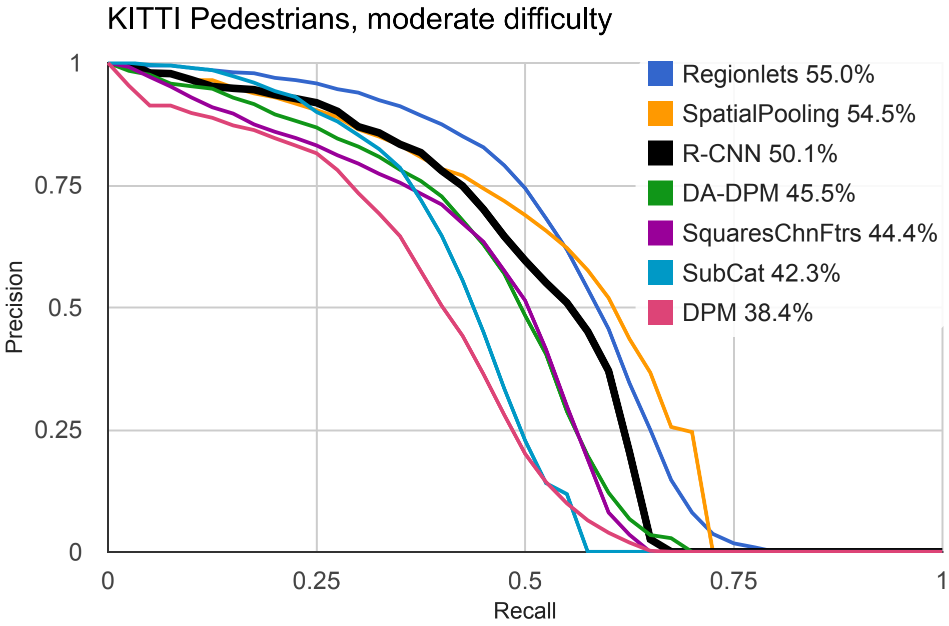


Figure 5.5: *AlexNet* results on the *KITTI* test set.

5.5.4 Error analysis

Results from the previous section are encouraging, but not as good as could be expected based on how *R-CNN* fares on *PASCAL VOC* compared to classical detectors. So what are the limits on performance? The proposals? The localisation accuracy of the CNN?

A cursory look at the highest scoring false positives suggests that the problem is localisation errors, made by the proposal method as well the *R-CNN* classifier and even errors present in the ground truth. By localisation, we mean predicting accurate bounding box coordinates once we've identified the presence of a pedestrian in some part of the image.

To quantify this effect we rerun the *Caltech* evaluation but remove all false positives that overlap with an annotation. This experiment provides an upper bound on performance that assumes precise localisation in detectors together with ideal non-maximum suppression. We see a surprisingly consistent and limited improvement for all methods of not more than 2% laMR. This means that our initial guess based on looking at false positives is wrong and actually almost all of the mistakes that worsen the laMR are actually background windows that are mistaken for pedestrians. What is striking about this result is that this is not just the case for our *R-CNN* experiments on detection proposals but also for methods that are trained as a sliding window detector. In the next chapter, we will see that fixing localisation errors tends to have a bigger effect in the low FPPI range not considered for evaluation (between $[10^{-4}, 10^{0}]$) and thus do not show up in the standard metrics.

Architecture training	# parameters	Test laMR Caltech1x	 Caltech10x
CifarNet	$\sim 10^5$	30.7%	28.4%
MediumNet	$\sim 10^6$	–	27.9%
AlexNet	$\sim 10^7$	32.4%	27.5%
SquaresChnFtrs (Chapter 4)		34.8%	

Table 5.10: Selection of results from previous sections when training different networks solely using *Caltech* training data. laMR: log-average miss rate on *Caltech* test set.

5.6 Small or big CNN?

So far we analysed *CifarNet* and *AlexNet* separately, and now compare them side by side. Tab. 5.10 shows performance on the *Caltech* test set for models that have been trained only on *Caltech*1× and *Caltech*10×. With a smaller training set *CifarNet* reaches 30.7% laMR, performing 2pp better than *AlexNet*. On *Caltech*10×, we find *CifarNet* performance improves to 28.4%, while *AlexNet* improves to 27.1% laMR. The trend confirms the intuition that lower capacity models saturate earlier when increasing the amount of training data than models with higher capacity. We also conclude that *AlexNet* would profit from better regularisation when training on *Caltech*1×.

Timing. Runtime during detection is ~3ms per proposal window. This is too slow for sliding window detection, but given a fast proposal method with high recall at fewer than 100 windows per image, scoring takes about 300ms per image. In our experience, *SquaresChnFtrs* runs in 2s per image, so the proposal stage is more expensive.

5.7 Takeaways

Work preceding this study suggested that CNNs for pedestrian detection underperform, despite sophisticated architectures with problem-specific modelling (see Chapters 2 and 4). In this chapter we showed that neither has to be the case. We present a wide range of experiments with two off-the-shelf models that reach competitive performance: the small *CifarNet* and the big *AlexNet*.

We present two networks that are trained on *Caltech* only, which outperform all previously published CNNs on *Caltech*. The *CifarNet* shows better performance than related work, even when using the same training data as the respective methods (Sec. 5.4.2). Despite its size, *AlexNet* also improves over previous CNNs even when it is trained on *Caltech* only (Sec. 5.5.2).

At time of publication we advanced the state of the art for pedestrian detectors that have been trained on *Caltech*1× and *Caltech*10×. The *CifarNet* was the best single-frame

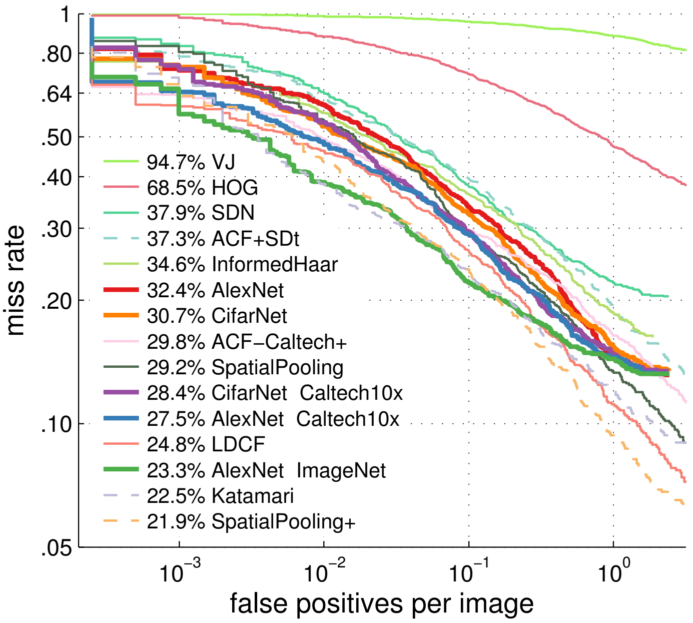


Figure 5.6: Comparison of our key results (thick lines) with published methods on *Caltech* test set. Methods using optical flow are dashed.

pedestrian detector that has been trained on *Caltech*1× (Sec. 5.4.2), while *AlexNet* was the best single-frame pedestrian detector trained on *Caltech*10× (Sec. 5.5.2).

In Fig. 5.6, we include all previously published methods on *Caltech* for the comparison, which also adds methods that use additional information at test time. *AlexNet* when pre-trained on *ImageNet* yields results that are competitive with the best previously published methods, but without using additional information at test time (Sec. 5.5.1).

We report first results for CNNs on the *KITTI* pedestrian detection benchmark. *AlexNet* improves over the proposal method (another pedestrian detector) but there is still room to further improve *KITTI* performance with CNNs.

5.8 Conclusion

We have presented extensive experimental evidence for the effectiveness of CNNs for pedestrian detection. Compared to previous CNNs applied to pedestrian detection our approach avoids problem-specific design. When using the exact same proposals and training data as previous approaches our “vanilla” networks outperform previous results.

We have shown that with pre-training on surrogate tasks, CNNs can reach top performance on this task. Interestingly we have shown that even without pre-training competitive results can be achieved, and this result is quite insensitive to the model size (from 10^5 to 10^7 parameters). Our experiments also detail which parameters are most critical to achieve top performance. At the time of publication of this study, we report the best known results for CNNs on both the challenging *Caltech* and *KITTI* datasets.

Our experience with CNNs indicates that they show good promise on pedestrian detection, and that reported best practices do transfer to this task. That being said, on this more mature field we do not yet observe the large improvement seen on datasets such as *PASCAL VOC* and *ImageNet*.

6 Towards a Human Baseline for Pedestrian Detection

With our retrospective analysis in Chapter 4, we demonstrated how crucial feature design has traditionally been for improving pedestrian detection performance. In Chapter 5, we showed that standard convolutional neural networks (CNNs) when trained appropriately work better than custom-designed networks with domain-specific elements. Since the publication of that work, more research on CNNs for pedestrian detection has resulted in significant improvements without signs of slowing down.

In this chapter, rather than looking back we want to look forward and characterise the gap between the current state of the art and the "perfect single frame detector". To this end, we set a human baseline for the *Caltech* dataset with a manual re-annotation of the test set in a detection-like setting. Furthermore, we manually group the errors of a top detector, allowing us to perform a more targeted analysis of performance, and try to separate out the impact of improving classification and localisation precision. Our results suggest shortcomings with the original annotations that may preclude a reliable evaluation especially when precise localisation is required.

To that end, we undertake a full reannotation of the training and test sets with a model in-the-loop to help cover the full training sequences. We demonstrate that our improved annotation protocol leads to better-aligned bounding boxes, which in turn results in higher quality detections when used for training. This also suggests that dealing with label noise in training remains an issue. We provide the sanitised set of training and test annotations for future research. We also suggest modifications to the standard *Caltech* evaluation metric to better reflect improvements on localisation performance.

Finally, based on our preliminary analysis and reannotations, we revisit *R-CNN*-like detectors consisting of a classical detector that supplies proposals and a CNN-based verification step (a lá Chapter 5). CNNs exhibit superior classification performance, but are less precise than the top classical detectors. Bounding box regression helps, but this suggests that improved architectures are necessary to make progress on addressing localisation errors.

An earlier version of this work was published at CVPR (Zhang *et al.*, 2016b) and subsequently at PAMI (Zhang *et al.*, 2018a). Shanshan Zhang was the lead author and provided most experiments, Rodrigo Benenson and Mohamed Omran provided annotations and contributed to the analysis and discussion, while Jan Hosang contributed the *AlexNet* experiments in Sec. 6.5.2 and the localisation/background analysis in Sec. 6.3.3 and Sec. 6.5.2.

6.1 Introduction

Despite the extensive research on pedestrian detection, recent methods still attain significant improvements, suggesting that a saturation point has not yet been reached. In this chapter we analyse the gap between the state of the art and a newly created human baseline (Sec. 6.3.1) for the *Caltech Pedestrian Dataset* (*Caltech*) (Dollár *et al.*, 2012b). The results indicate that there is still a ten-fold improvement to be made before reaching human-like performance on this benchmark. We aim to investigate which factors will help close this gap.

We analyse the errors made by top performing pedestrian detectors and make recommendations for addressing these. We conduct several kinds of analyses (Sec. 6.3.2), including manual inspection, automated analysis of certain problem cases (e.g. blur, contrast), and oracle experiments to isolate sources of error. Our results indicate that inaccurate localisation is an important source of high-confidence false positives. We address this by improving the training set alignment quality, both by manually sanitising the default *Caltech* training annotations, as well as fixing the remaining annotations algorithmically (Sec. 6.4 and Sec. 6.5.1).

To address problems with foreground-background discrimination, we study CNNs for pedestrian detection given their strength at object classification (Chapter 5, Krizhevsky *et al.* 2012) and discuss which factors affect their performance (Sec. 6.5.2).

6.1.1 Contributions

Our key contributions are as follows:

1. We analyse the performance of a state-of-the-art pedestrian detector, providing detailed insights into its shortcomings.

2. We set a human baseline for the *Caltech* benchmark; and extend the resulting detections to a full, sanitised version of the annotations. These can serve as new, high quality ground truth for training and test sets and are publicly available[6].

3. We analyse the effects of training data quality and determine the effect of better aligned and labelled annotations on performance.

4. Based on the above, we explore variants of top-performing methods: the *Filtered Channel Features* (*Checkerboards*) detector of Zhang *et al.* (2015b) and *R-CNN*

[6] http://www.mpi-inf.mpg.de/pedestrian_detection_cvpr16

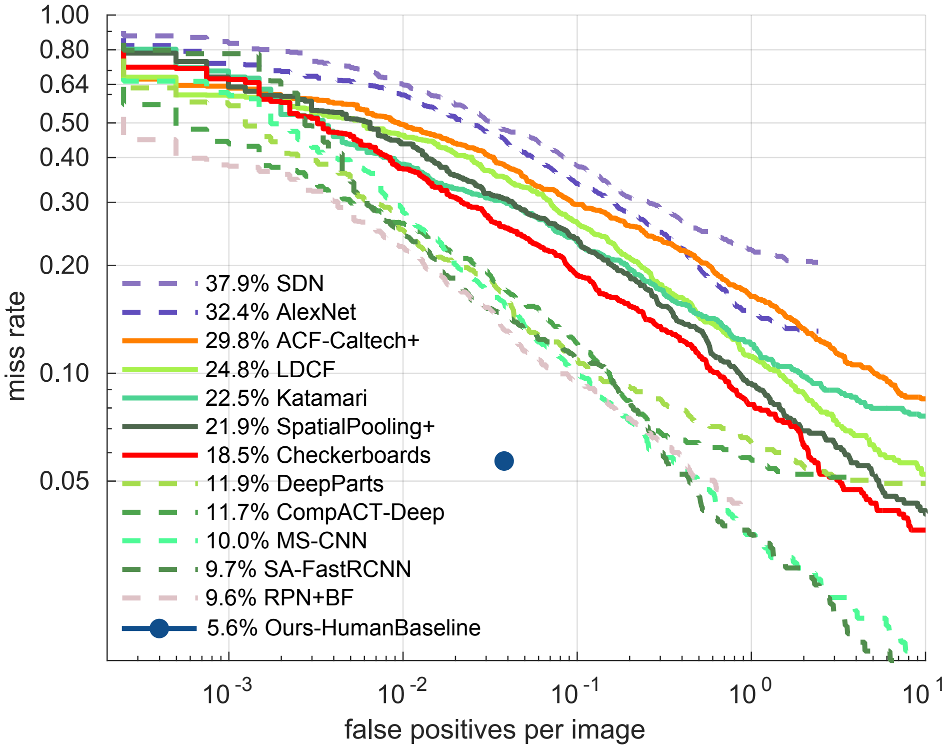


Figure 6.1: Overview of the top results on the *Caltech* benchmark. At ∼95% recall, state-of-the-art detectors make ten times more errors than the human baseline.

(Girshick *et al.* 2014 and Chapter 5), and demonstrate improvements over the baselines.

6.2 Preliminaries

Before presenting our analysis, we want to introduce the experimental setting, including the relevant datasets, evaluation metrics and baseline detectors. We conduct our analysis in this chapter on the *Caltech* (Dollár *et al.*, 2012b) and *KITTI* (Geiger *et al.*, 2012) datasets, which we describe at length in Sec. 2.2. In the discussion that follows, we distinguish between classification and localisation. Given an image window, the detector needs to classify it as either corresponding to background or pedestrian. In the latter case, the detector needs to additionally localise the pedestrian, i.e. produce accurate bounding box coordinates. These are not independent tasks but useful to consider separately for the purpose of our analysis.

Filter type	MR^O_{-2}
ACF (Dollár *et al.*, 2014)	44.2
SquaresChnFtrs (Benenson *et al.*, 2014)	34.8
LDCF (Nam *et al.*, 2014)	24.8
RotatedFilters	19.2
Checkerboards (Zhang *et al.*, 2015b)	18.5

Table 6.1: The type of filters applied to the baseline feature channels strongly determines the performance of detectors in the *ICF* family.

Base detector	MR^O_{-2}	+Context $\Delta\mathrm{MR}^O_{-2}$	+Flow $\Delta\mathrm{MR}^O_{-2}$
Orig. 2Ped (Ouyang and Wang, 2013b)	48	+ 5	/
Orig. SDt (Park *et al.*, 2013)	45	/	+ 8
SquaresChnFtrs (Chapter 4)	35	+ 5	+ 4
Checkerboards (Zhang *et al.*, 2015b)	19	+ 0	+ 1

Table 6.2: Performance improvement as a result of adding context (Ouyang and Wang, 2013b) or optical flow (Park *et al.*, 2013) to different baseline detectors.

6.2.1 Evaluation metrics

***MR_{-2}*, *MR_{-4}*.** For the standard *Caltech* evaluation procedure (Dollár *et al.* 2012b, Sec. 2.1), the miss rate is averaged over the low precision range of $[10^{-2}, 10^0]$ FPPI (false positives per image). In the course of our analysis, we found that this does not adequately reflect improved localisation performance. The latter instead affects the lowest FPPI rates. Accordingly, we extend the support region of the log-average operation from the standard range of $[10^{-2}, 10^0]$ to an expanded range of $[10^{-4}, 10^0]$. We will refer to these metrics as MR_{-2} and MR_{-4} respectively, and drop the reference to the "log-average" in the abbreviation for the sake of readability. We expect the MR_{-4} metric to become more important as detectors get stronger.

MR^O, MR^N. In Sec. 6.4 we introduce new annotations for the test set. We show evaluations on both original and new annotations for a more comprehensive (and backward-compatible) comparison. The O superscript indicates the use of the original annotations for evaluation, and N the use of the new ones.

In total, we thus use four evaluation metrics: MR^N_{-2} , MR^N_{-4} , MR^O_{-2} and MR^O_{-4} for our *Caltech* experiments in this chapter.

6.2.2 Detectors: Filtered Channel Features

We consider two members of the *ICF* detector family (Dollár *et al.*, 2009a) for our analysis. The *Checkerboards* detector of (Zhang *et al.*, 2015b) is the top detector in this family on the *Caltech* benchmark at the time of writing. It is an extension of the original *ICF* detector, which applies various filters to the base HOG+LUV feature channels before feeding them to a boosted decision forest for classification. Additionally, we also consider the *RotatedFilters* detector, which is a simplified variant of *LDCF* (Nam *et al.*, 2014), that almost matches the performance of *Checkerboards* while being 6× faster at training and test time.

We compare the performance of several detectors from the *ICF* family in Tab. 6.1, where we can see a big improvement from 44.2% to 18.5% MR^{O}_{-2} by introducing filters over the feature channels and optimising the filter bank. These results complement the experiment in Chapter 4 (Sec. 4.4.1), which demonstrates how merely improving features while keeping the method otherwise fixed reproduces performance improvements achieved in a decade of work on pedestrian detection. There, we mostly varied the base feature channels (from a simple luminance channel to 10 HOG+LUV channels), whereas here we show the effect of additionally applying filters to these channels.

We should also note that many top-performing CNN-based methods (Fig. 6.1) use *ICF* detectors for generating pedestrian hypotheses, e.g. *DeepParts* (Tian *et al.*, 2015a), *CompACT-Deep* (Cai *et al.*, 2015), and *SA-FastRCNN* (Li *et al.*, 2018). Therefore, the insights derived from analysing *RotatedFilters* and *Checkerboards* are also applicable to other top methods.

Additional cues. The review in Chapter 4 showed that context and optical flow information can help improve detections. However, as the detector quality improves (Tab. 6.1) the benefit of including such additional cues erodes (Tab. 6.2). It is plausible that with some adaptation more gains can be squeezed out from the use of these cues, but for the purposes of our current analysis we will only consider pure *ICF* detectors.

6.2.3 Detectors: CNN-based

In the standard *R-CNN* framework (Girshick *et al.*, 2014), external methods are used to generate detection proposals, which are then fed into CNNs for feature extraction and classification. Such a two-stage strategy saves computation by reducing the number of windows for CNNs to process, but on the other hand introduces a dependency of the final detection results on proposal quality — especially in terms of recall. Tweaking the sensitivity of the proposal method to generate more hypotheses can thus be helpful to potentially achieve higher recall, but also increases the chance for false positives.

As reported in Chapter 5 and Tian *et al.* (2015b), current top-performing CNN methods are sensitive to the underlying detection proposals. We thus first focus on improving the proposals by optimising the *Filtered Channel Features* detectors, and turn to the CNNs themselves in Sec. 6.5.2).

6.3 Preliminary Analysis of the State of the Art

In this section we seek to understand the shortcomings of current detectors. To this end, we estimate a lower bound for the log-average miss rate on *Caltech* with a human baseline detector. We then manually categorise the mistakes of state-of-the-art detectors, identifying several problem areas. Besides the fine-grained categorisation, errors can be roughly separated into localisation and classification errors. With the help of an oracle, we identify the contribution of each to the performance of several recent detectors. Our results all in all suggest that the current *Caltech* annotations preclude a precise evaluation and might be holding current methods back, which we address in later sections.

6.3.1 Are we reaching saturation?

How much progress can still be expected on current pedestrian detection benchmarks? To answer this question, we propose to use a human baseline as a lower bound: As domain experts familiar with the benchmarks, we manually "detect" pedestrians in the *Caltech* test set with an improved annotation protocol.

Human baseline protocol. To collect human detections, we used custom annotation software. We discuss two important design decisions: the process of marking a pedestrian, as well as the order of presenting frames.

The *Caltech* benchmark normalises the aspect ratio of all bounding boxes to 0.41 (Dollár *et al.*, 2012b). Since the original annotations cover the extent of the full body, this can result in inconsistent alignments especially for asymmetrical poses (e.g. if an arm or a leg is outstretched). To remedy this, we resorted to a different annotation scheme than the usual one: Rather than marking upper and lower left corners of the bounding box, we annotated pedestrians by drawing a line the top of the head to a point between both feet. A bounding box is then automatically generated such that its centre coincides with the centre point of the manually-drawn axis (see Fig. 6.2). This procedure ensures that the box is well-centred on the subject, which is hard to achieve when marking the bounding box corners.

To ensure a fair comparison with existing detectors, most of which operate over a single frame at a time, we focus on the single-frame monocular detection setting. Our software thus presents frames in random order and without access to surrounding frames

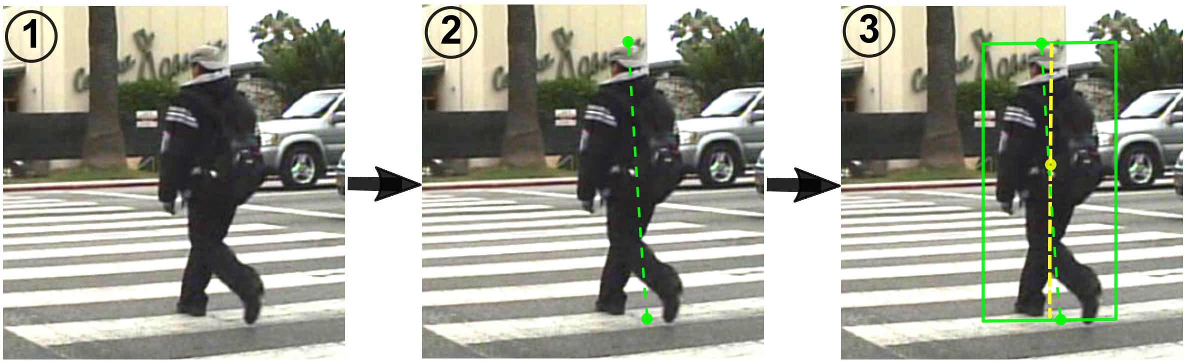


Figure 6.2: Illustration of bounding box generation for human baseline. The annotator only needs to draw a line from the top of the head to the central point between both feet, a tight bounding box is then automatically generated with the desired aspect ratio.

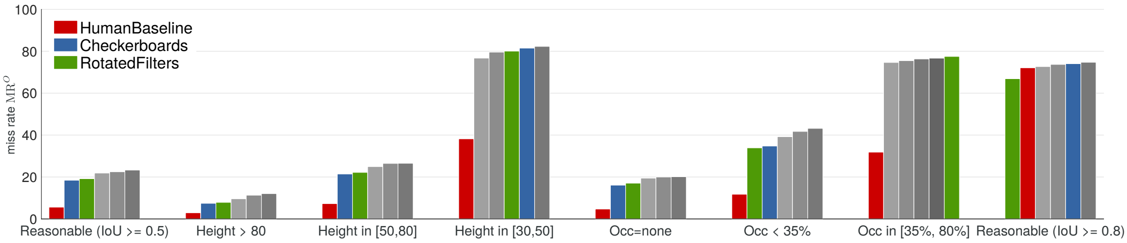


Figure 6.3: Detection quality (log-average miss rate) for different subsets of the *Caltech* test set. Each group shows the human baseline, the *Checkerboards* (Zhang *et al.*, 2015b) and *RotatedFilters* detectors (see legend), as well as three other highest-ranked methods (different for each setting).

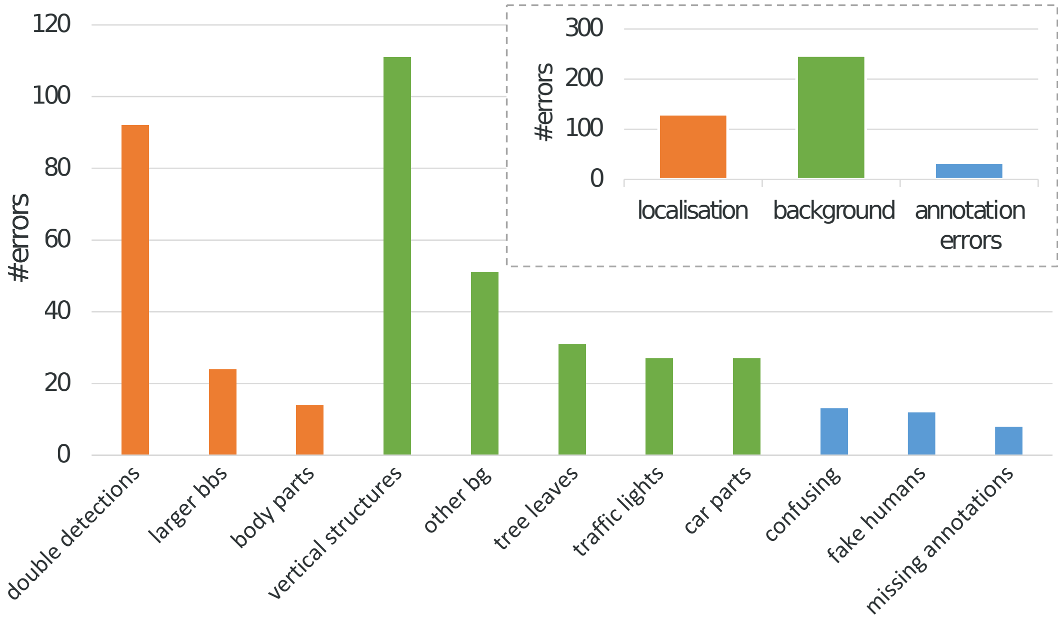
from the source videos. While this does not control for annotators remembering their decisions for past frames, this still helps ensure that detections are mostly dependent on appearance and single-frame context rather than long-term motion.

To check for consistency among the two annotators who generated these results, we let both annotate a subset of test images ($\sim 10\%$) and evaluated these separately. With an Intersection over Union (IoU) ≥ 0.8 matching criterion, the results were identical up to a single bounding box.

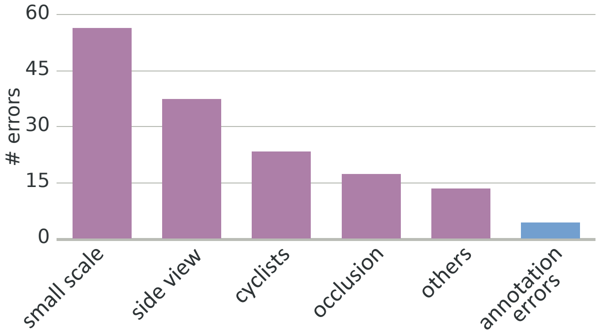
In Fig. 6.3, we compare our human baseline to *Checkerboards, RotatedFilters* and other competing methods on various subsets of the test data. We find that the human baseline outperforms state-of-the-art detectors under all settings[7]. We also notice the gap between human baseline and state-of-the-art detectors is especially large for harder cases, e.g. small-scale and heavily occluded pedestrians.

Fig. 6.3 also shows that *Checkerboards* and *RotatedFilters* perform well across all subsets. In the few cases where they are not top-ranked, all methods exhibit low detection quality. *Checkerboards* is not optimised for the most common case on the

[7] Except for IoU ≥ 0.8. This is due to inaccuracies of the ground truth, discussed in Sec. 6.4.



(a) Types of false positive errors



(b) Types of false negative errors

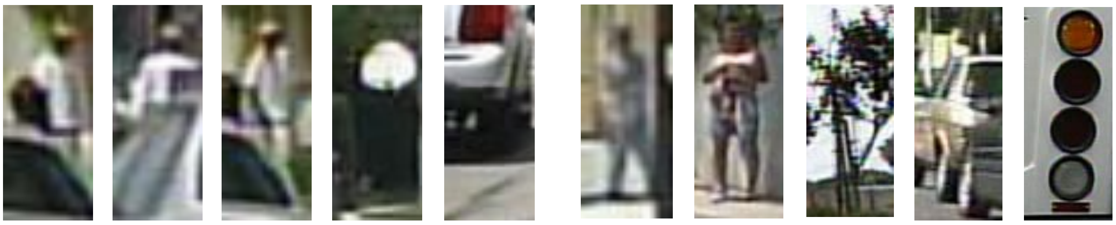
Figure 6.4: Error sources of *Checkerboards* (Zhang *et al.*, 2015b) on the *Caltech* test set.

Caltech dataset, but nevertheless shows good performance across a variety of situations and is thus an interesting method to analyse.

Conclusion. There is still room for improvement for automatic methods on the *Caltech* benchmark, and we have not yet reached saturation.

6.3.2 Manual Error Analysis

Since there is room for improvement for existing detectors: When and how do they currently fail? In this section we analyse the errors made by the *Checkerboards* detector, which obtains top performance on most subsets of the test set (Fig. 6.3). Since most



(a) Low-scoring objects

(b) High-scoring objects

Figure 6.5: Failure cases of *Checkerboards* (Zhang *et al.*, 2015b). Each group shows image patches of similar scores: some background objects have high scores, while some persons have low scores. We aim to understand when the detector fails through analysis.

top methods are from the *ICF* family (Fig. 6.1), we expect this analysis to apply more broadly. Methods that apply CNNs to proposals from *ICF* detectors are also affected.

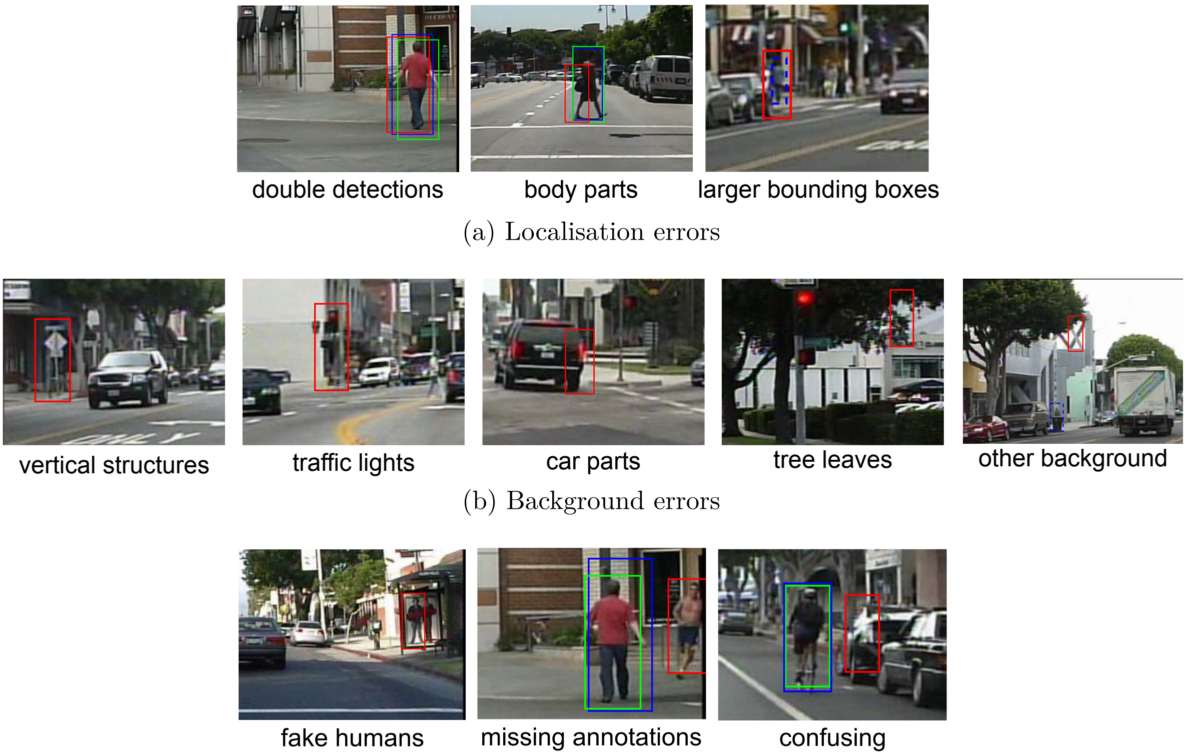
There are two types of errors a detector can make: false positives (detections on background, redundant detections, or poorly-localised detections) and false negatives (low-scoring or missing detections). Fig. 6.5 shows examples of both low-scoring and high-scoring detections, and each group contains a mixture of both error types.

In this analysis, we look at false positives and negatives at 0.1 FPPI, and manually cluster them into visually distinctive groups. A total of 402 false positives and 148 false negatives are categorised by error type, as shown in Fig. 6.4.

False positives. We manually assign false positives to one of eleven categories, shown in Fig. 6.4a. These fall into three groups: localisation, background, and annotation errors. We show examples for each category in Fig. 6.6. Localisation errors are defined as detections that have insufficient overlap with ground truth bounding boxes, whereas background errors are detections that don't overlap with any annotations at all (Figs. 6.6a to 6.6c).

Background errors are most the common type of false positive, and mainly correspond to vertical structures (e.g. Fig. 6.6b), and to a lesser to extent to other types of objects such as tree leaves and traffic lights. This indicates that the detectors could benefit from the inclusion of more *vertical context*, providing visibility over larger structures and a rough height estimate. In Sec. 6.5.2 we explore how to better handle background errors by using CNNs, which has a larger receptive field than *Checkerboards*, i.e. takes more context into account.

Localisation errors are dominated by double detections, i.e. high scoring detections covering the same person (see the first two examples in Fig. 6.6a). This indicates that improved detectors need to have more localised responses (peakier score maps) and/or a different non-maximum suppression strategy. In Sec. 6.4 and Sec. 6.5.1 we explore how to improve the detector localisation.



(a) Localisation errors

(b) Background errors

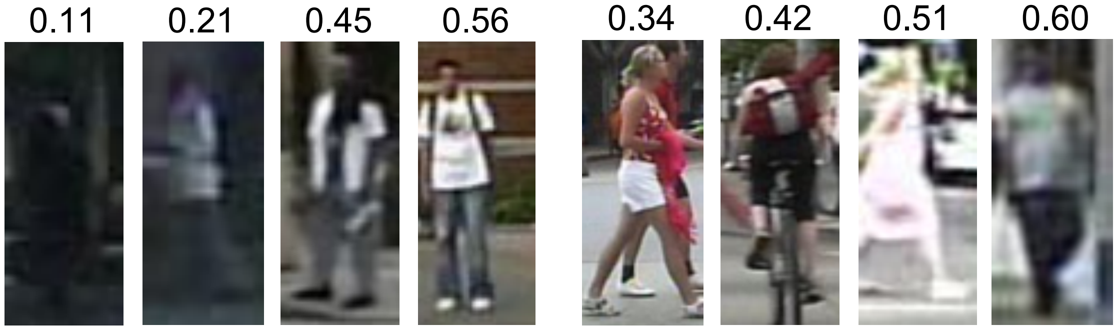
(c) Annotation errors

Figure 6.6: Different types of false positive errors made by *Checkerboards*. True/false positives in red/green, annotations in blue, and ignore regions in dashed blue.

Some detection errors can be traced back to problems with the annotations. These are mainly missing ignore regions, e.g. annotations that would otherwise exclude depictions of persons in the environment such as mannequins or billboards. There are also a handful of unmarked pedestrians. In Sec. 6.4 we revisit the *Caltech* annotations.

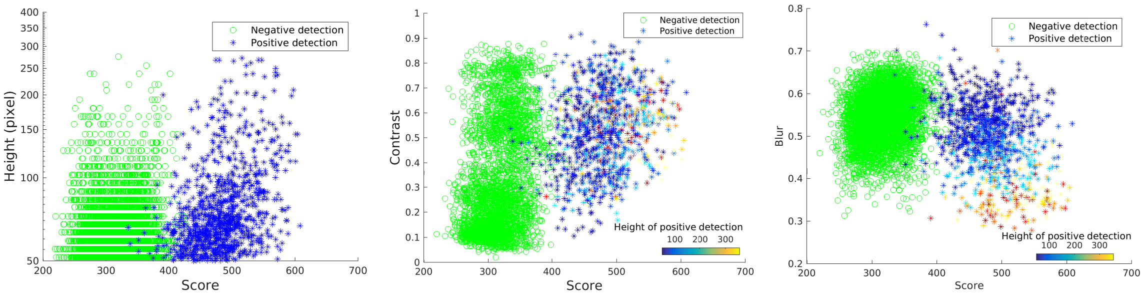
False negatives. Our clustering results in Fig. 6.4b reflect the well-established difficulty of detecting small and occluded objects. We hypothesise that poor performance on cyclists and persons viewed from the side may be the result of underrepresentation in the training set: Most persons are walking on the pavement with a trajectory parallel to that of the vehicle. Augmenting the training set with external images that address this bias of the dataset might be an effective strategy.

Is it scale? Or rather visual quality? For false negatives, a major source of errors is small scale. We additionally observe that small persons are commonly saturated (over- or under-exposed) and blurry. We thus hypothesise that this might also interfere with detection quality, besides the mere availability of fewer pixels for making a decision. To this end, we define two automated measures for contrast and blur that we apply to detections whether true or false. Contrast is measured via the difference between the



(a) Contrast (b) Blur

Figure 6.7: Examples for images with different levels of contrast/blur. The number on top of each image indicates the contrast/blur measure.



(a) Size versus score (b) Contrast versus score (c) Blur versus score

Figure 6.8: Correlation between size/contrast/blur and score.

top and bottom quantiles of the grey scale intensity of the pedestrian patch. Blur is measured as the difference between the input and its blurred patch, which is generated by applying a mean filter to input image (Crete *et al.*, 2007). Note that all patches are rescaled to the input size expected by our model (120 × 60 pixels) prior to measuring the degree of blur. Both contrast and blur measures have a range of [0, 1], and higher values indicate a higher degree of contrast or blur. Fig. 6.7 shows pedestrians ranked by our contrast and blur measures. One can observe that our quantitative measures correlate well with human notions of blur and contrast.

In order to investigate the three factors separately, we observe the correlation between size/contrast/blur and detection score, as shown in Fig. 6.8. We can see that the overlap between false positive and true positive is equally distributed across different levels of contrast and blur, while for scale, the overlap is quite high at small scale. Thus we conclude that small scale itself is the main factor negatively impacting detection quality and that high blur and low contrast are not.

Discussion. As a sanity check, we conduct the same analysis of errors for other detectors and datasets, namely for *Checkerboards* on *KITTI* and for *RPN+BF* (Zhang *et al.*,

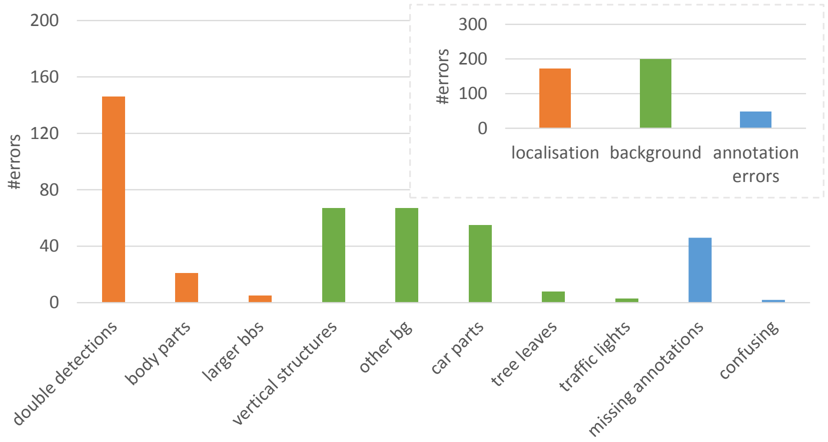


Figure 6.9: Types of false positive errors made by *Checkerboards* (Zhang *et al.*, 2015b) on the *KITTI* validation set.

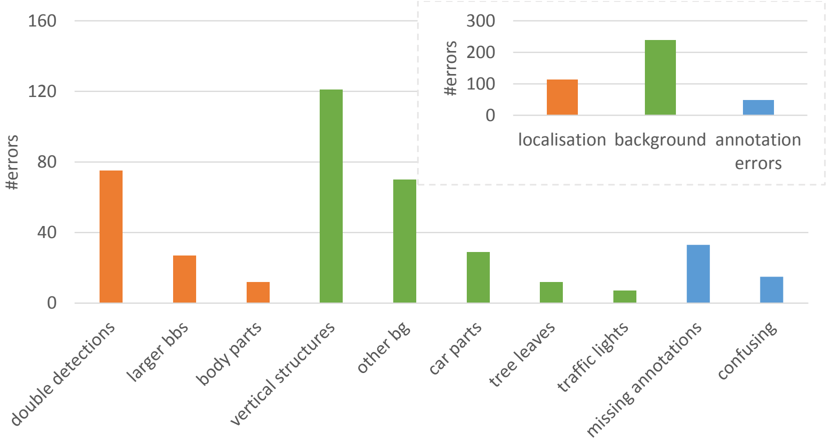
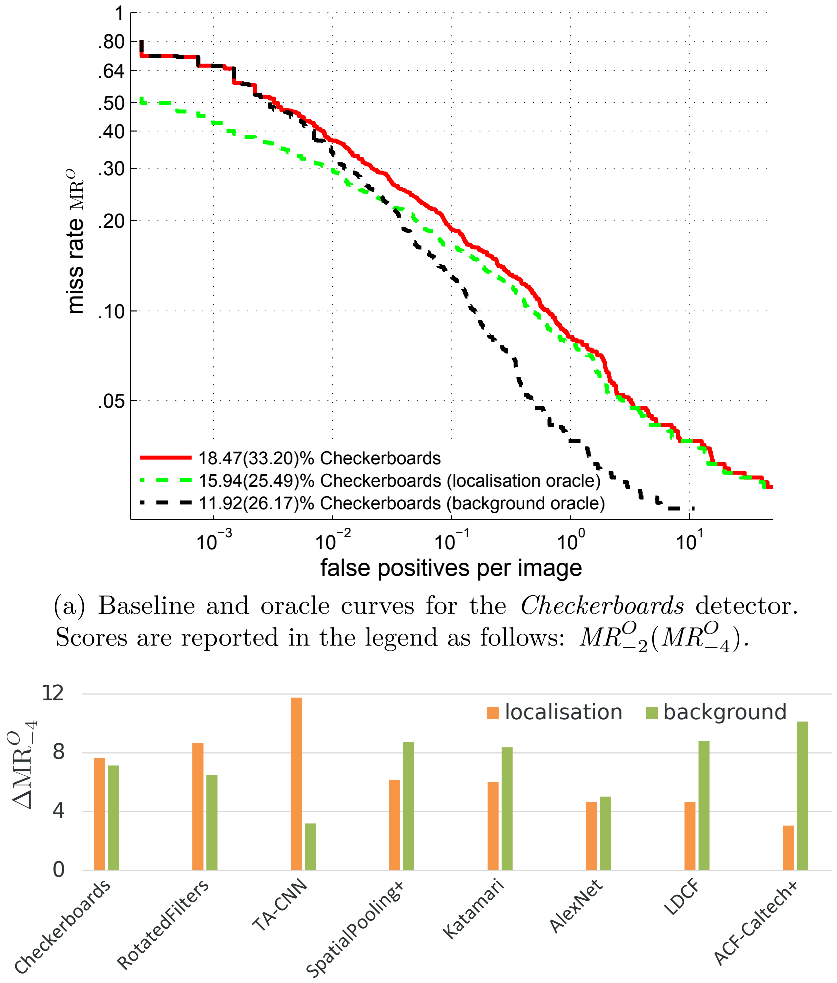


Figure 6.10: Types of false positive errors made by *RPN+BF* on the *Caltech* test set.

2016a), another state-of-the-art detector, on *Caltech*. While comparing the statistics shown in Figs. 6.4a, 6.9 and 6.10, we observe similar trends for the error sources, e.g. double detections, vertical structures, annotation errors.

Conclusion. Our analysis shows that false positive errors have well defined sources that can be specifically targeted with the strategies suggested above. A fraction of the false negatives are also addressable, although the small and occluded pedestrians remain a hard and significant problem.



(a) Baseline and oracle curves for the *Checkerboards* detector. Scores are reported in the legend as follows: $MR^O_{-2}(MR^O_{-4})$.

(b) Comparison of miss-rate gain ($\Delta\mathrm{MR}^O_{-4}$) for top performing methods.

Figure 6.11: Oracle cases evaluation over *Caltech* test set.

6.3.3 Oracle Experiments

So far, our analysis focused on error counts. For metrics that depend on the area under a performance curve (e.g. log-average miss rate or average precision), high-scoring errors matter more than low-scoring ones as they more strongly impact the curve's trajectory: The miss rate is computed from fewer detections early on. In this section we use oracle test cases to directly measure the impact of two types of errors on *Caltech*: localisation errors and positive classification errors (i.e. misidentifying background as foreground). In the oracle case for localisation, all false positives that overlap with ground truth are ignored for evaluation. In the oracle tests for positive classification, all false positives that do not overlap with ground truth are ignored.

Fig. 6.11a shows that fixing localisation mistakes improves performance in the low FPPI region, while fixing background mistakes improves results in the high FPPI region. In Fig. 6.11b we show the gains that can be obtained in terms of MR^O_{-4} by fixing localisation or positive classification issues. When comparing the eight top-performing methods we find that fixing either problem would boost performance significantly for most. It is important to note that localisation and positive classification errors together comprise all false positives. If we were to remove both types, the only mistakes that would remain are missed pedestrians and the result would be a horizontal line with very low miss rate. However, due to the log-scale, the sum of localisation and background deltas do not add up to the total miss rate.

Conclusion. For most top performing methods localisation and positive classification errors are an issue, however CNN-based methods are less affected by the latter.

6.4 Reannotating Caltech

When evaluating our human baseline and other methods with a strict criterion of IoU ≥ 0.8, we notice that performance drops for all methods (Fig. 6.3), and — perhaps counter-intuitively — our human baseline no longer outperforms the rest. Our analysis also shows that localisation errors are a problem for most detectors, and even for CNN-based detectors which make fewer classification errors than the rest. Taken together, these results suggest that the *Caltech* annotations might get in the way of a reliable evaluation, especially for closing the final gap on this benchmark.

The original annotation protocol is based on interpolating sparse annotations across multiple frames (Dollár *et al.*, 2012b), and these annotations are not necessarily located on the evaluated frames. Upon inspection we notice that this interpolation indeed generates a systematic spatial offset in the annotations. Humans walk with a natural vertical oscillation that is not captured by the linear interpolation scheme. This effect is not noticeable when using a generous IoU ≥ 0.5 threshold, but causes problems when we require more precise localisation.

To fix this issue together with the errors previously identified, we create a new set of improved annotations for *Caltech*. Our aim is two-fold: On the one hand, we want to provide a more accurate evaluation of the state of the art, especially in light of low log-average miss rates on the "Reasonable" set. On the other hand, with the high quality

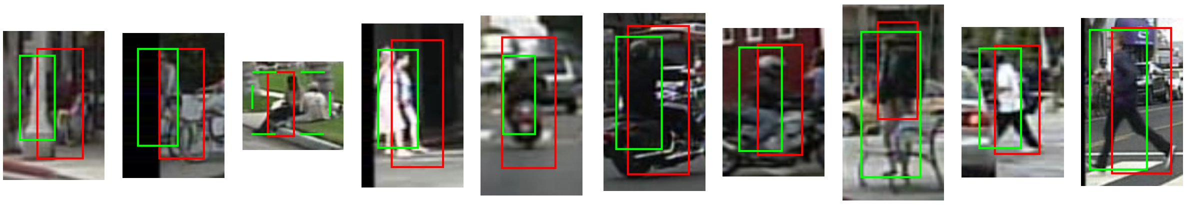
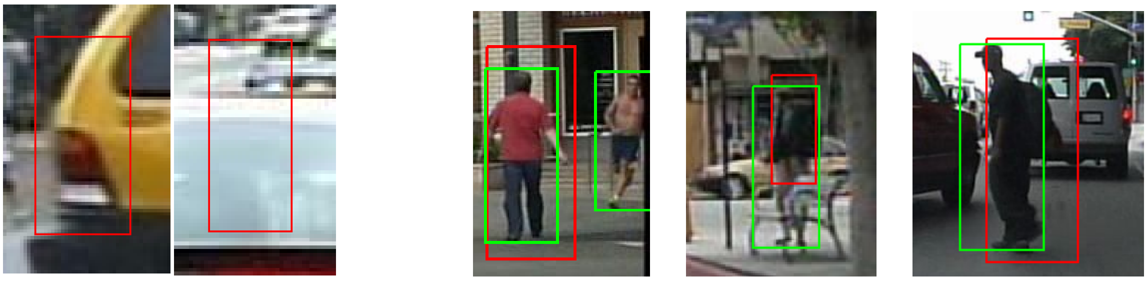


Figure 6.12: Examples of original (red) vs. new annotations (green). Ignore regions are marked with dashed lines. These are the ten annotation pairs with the largest IoU gap.



(a) False annotations

(b) Poor alignment

Figure 6.13: Examples of errors in original annotations. New annotations in green, original ones in red.

training annotations we could evaluate how much these lead to better detections — or in other words, how sensitive detectors are to label noise (Sec. 6.5.1).

6.4.1 Manual Single-frame Annotation Protocol

We re-annotate both the *Caltech*1× training and test sets (Sec. 2.2), and focus on high quality. We use the same labelling procedure as for our human baseline but with some modifications. As before, each person is annotated with a line from the top of the head to the point between both feet. The annotators must hallucinate head and feet if these are not visible. However, when the person is not fully visible, they must now also annotate a rectangle around the largest visible region. This allows us to estimate the occlusion level as with the original annotations. Additionally, annotators are allowed to look at the full video to decide if a person is present or not. They are requested to mark ignore regions in areas covering crowds, human shapes that are not persons (posters, statues, etc.), and in areas where the presence of pedestrians could not be excluded with certainty. After creating a full independent set of annotations, we validated the new annotations against the originals. We added any correct annotation from the original set that was not accounted for in the new set.

In summary, our new annotations differ from the human baseline in the following aspects: both training and test sets are annotated, ignore regions and occlusions are also annotated, the video data is used for decisions, and multiple revisions of the same image are allowed.

We show some examples of differences between original and new annotations in Fig. 6.12. Our new annotations correct several types of errors in the existing annotations, such as misalignments (Fig. 6.13b), missing annotations (false negatives), false annotations (false positives, Fig. 6.13a), and the inconsistent use of “ignore” regions.

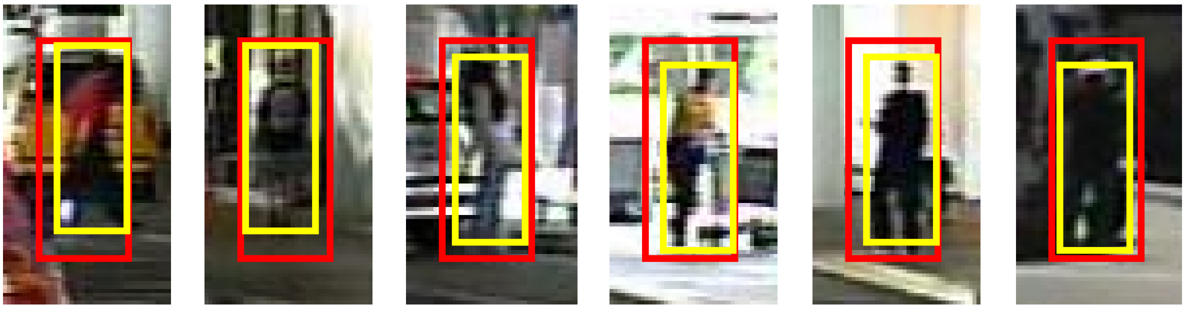


Figure 6.14: Examples of automatically realigned ground truth annotations. Red/yellow→ before/after realignment.

1×	10× data	MR^O_{-2} (MR^O_{-4})	MR^N_{-2} (MR^N_{-4})
data	aligned with		
Orig.	Ø	19.20 (34.28)	17.22 (31.65)
Orig.	Orig. 10×	19.16 (32.28)	15.71 (28.13)
Orig.	New $^1/_2$×	16.97 (28.01)	14.54 (25.06)
New	New 1×	16.77 (29.76)	12.96 (22.20)

Table 6.3: Test set performance of *RotatedFilters* when using training annotations re-aligned in different ways. All models trained with *Caltech*10×, composed with different $1\times + 9\times$ combinations.

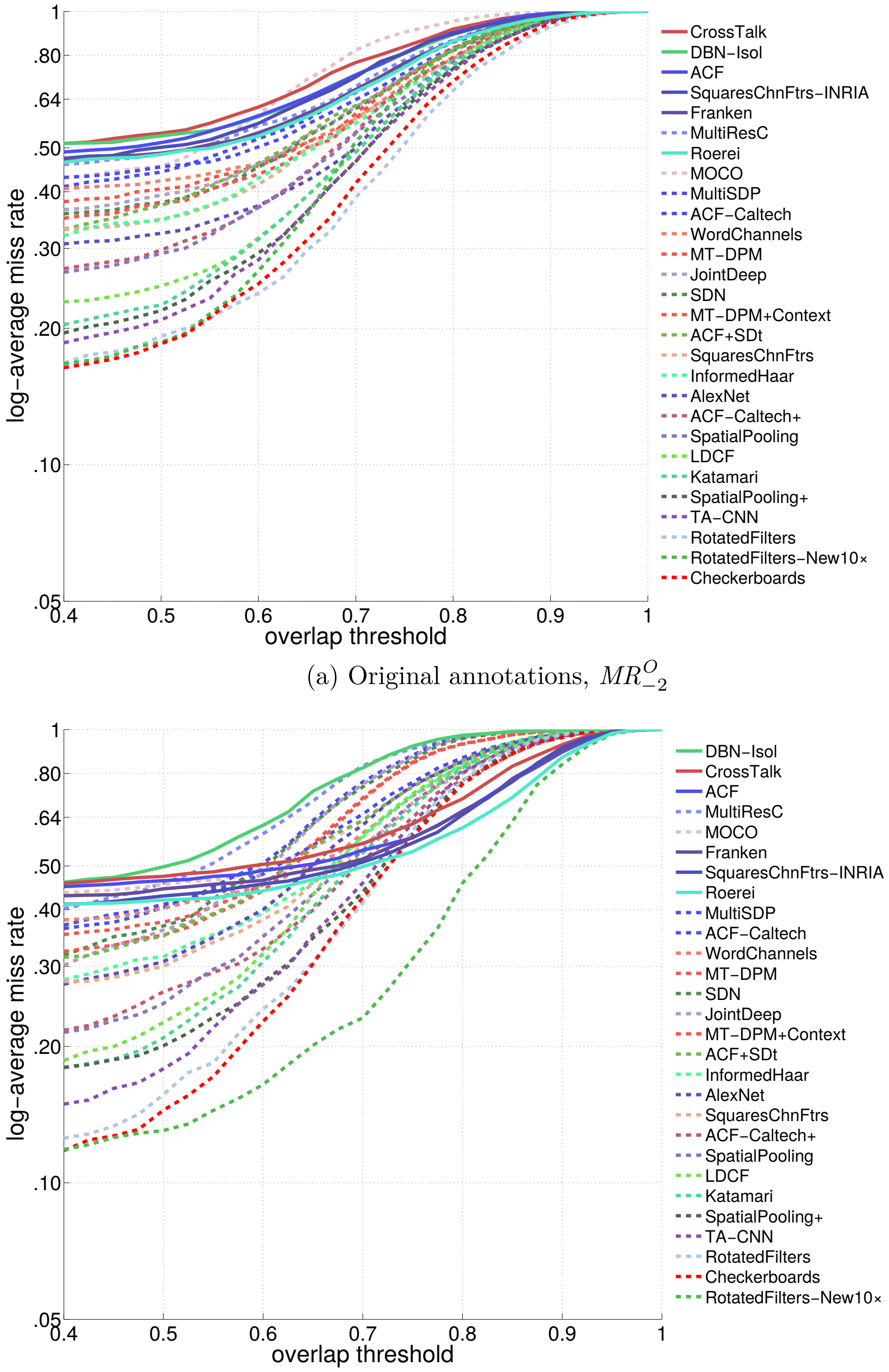
6.4.2 Semi-automatic Sequence Annotation

The detectors we consider here, whether from the *ICF* family (Nam *et al.*, 2014; Zhang *et al.*, 2015b) or CNN-based (Chapter 5), benefit from an expanded training set: *Caltech*10× vs. *Caltech*1×. Since we only manually reannotate the *Caltech*1× images, we use a model trained on these new annotations to re-align the annotations in the remaining frames. Fig. 6.14 shows example results of this process.

In Tab. 6.3 we report results for different sets of training annotations and different automatic realignment schemes. The results indicate that using a detector to improve overall data alignment is indeed effective across both metrics, especially when the realignment model is trained with the more accurate annotations — even with a partial set ($^1/_2$). This is in line with the analysis of Sec. 6.3.2.

6.4.3 Quantitative Analysis

Prior to using the new annotations for training in the next section, we examine their impact when evaluating methods trained on the original annotations.



(a) Original annotations, MR^{O}_{-2}

(b) New annotations, MR^{N}_{-2}

Figure 6.15: Plot of log-average miss rate versus overlap threshold (IoU) for the top-performing methods on the "Reasonable" experimental setting. When evaluating against the new annotations, methods trained on *INRIA* (represented with solid curves) are better behaved than methods trained on the original *Caltech* annotations when stricter overlap criteria are applied.

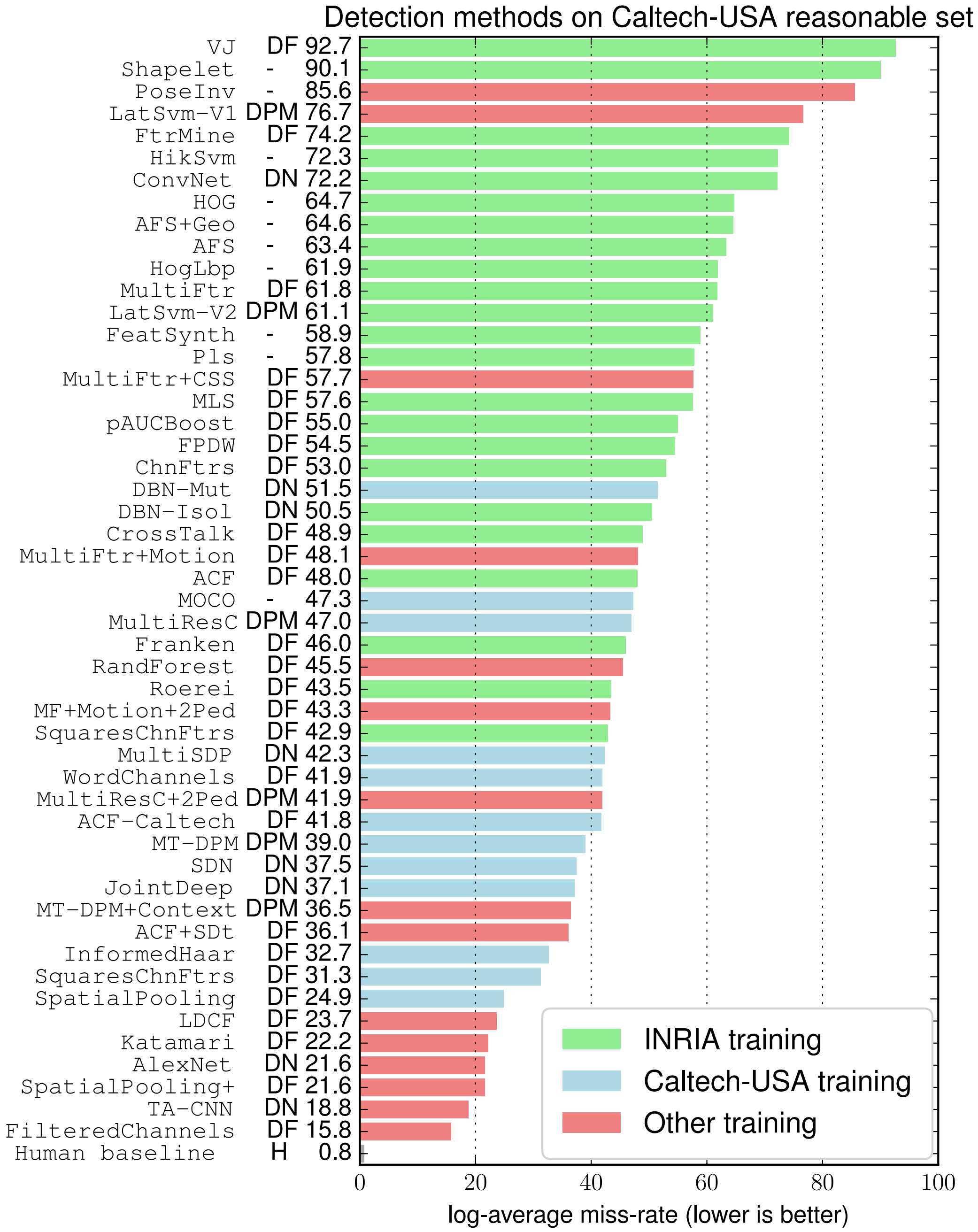


Figure 6.16: Ranking of methods when evaluated against the new *Caltech* annotations on the “Reasonable” test set (MR_{-2}^{N}). As in Chapter 4, `DF`: decision forest, `DPM`: deformable parts model, `DN`: deep network.

Alignment Quality. Fig. 6.15 plots $\mathrm{MR}^{\mathrm{O}}_{-2}$ and $\mathrm{MR}^{\mathrm{N}}_{-2}$ of top performing methods versus the IoU criterion for accepting detections as true positives. The standard evaluation uses a threshold of 0.5. On these plots, methods trained on *INRIA* have continuous lines, methods trained on *Caltech* dashed ones (see also Fig. 6.16).

In Fig. 6.15a (original annotations) the ranking of the methods remains stable as the overlap threshold becomes stricter, which is consistent with observations in Dollár *et al.* (2012b). Interestingly, we observe a different trend in Fig. 6.15b, where all methods are evaluated against the new annotations ($\mathrm{MR}^{\mathrm{N}}_{-2}$). Those methods trained on *INRIA*, while performing poorly at $\mathrm{IoU} = 0.5$, perform comparatively well at higher IoU thresholds, eventually passing all methods trained on the original *Caltech* data. We attribute this to the fact that *INRIA* annotations are of higher quality (esp. in terms of alignment), which is reflected in detector localisation ability.

This discrepancy between original and new annotations confirms that our improved annotations are better with respect to localisation.

Re-ranking the State-of-the-Art. As a sanity check, we re-rank all published *Caltech* results (Fig. 6.16) using the new annotations ($\mathrm{MR}^{\mathrm{N}}_{-2}$). Compared to the MR^{O}_{-2} metric, the overall trend is preserved — some minor ranking changes notwithstanding (e.g. *JointDeep* versus *SDN*). This is a good sign that the improved annotations are not a radical departure from previous ones. As discussed previously, the improved annotations matter most for future methods that aim to make progress on the hardest cases, especially in the low FPPI region where high-confidence mistakes show up.

6.5 Improving the state of the art

In the previous section, we described our reannotation protocol for *Caltech*. We provided evidence that these are indeed better aligned. The behaviour of the methods trained on INRIA with stricter IoU-thresholds for evaluation suggests that well-aligned labels are critical for good localisation performance. Equipped with the new training annotations, we can examine this further and measure their impact on both localisation and overall detection quality.

6.5.1 Impact of Improved Training Annotations

Examining localisation quality. In Tab. 6.4 we measure a detector's localisation quality via the median IoU between true positive detections and a given set of annotations. When evaluating with the original annotations ("Median IoU^{O}" column), only the model trained with the original annotations has good localisation. However, when evaluating with the new annotations ("Median IoU^{N}" column) *both* models trained either on *INRIA* or with the new annotations reach high localisation accuracy. This indicates that our

Detector	Training data	Median IoU^O	Median IoU^N
Roerei (Benenson *et al.*, 2013)	INRIA	0.76	*0.84*
RotatedFilters	Orig. 10×	*0.80*	0.77
RotatedFilters	New 10×	0.76	*0.85*

Table 6.4: Median IoU of true positives for detectors trained on different data, evaluated on both original and new *Caltech* annotations. Models trained on *INRIA* align well with our new annotations, confirming that they are more precise than previous ones.

Detector	Anno. variant	MR^O_{-2}	MR^N_{-2}
ACF (Dollár *et al.*, 2014)	Original	*36.90*	40.97
	Pruned	36.41	35.62
	New	41.29	*34.33*
RotatedFilters	Original	*28.63*	33.03
	Pruned	23.87	25.91
	New	31.65	*25.74*

Table 6.5: Effects of different training annotation sets on detection quality on validation set performance (*Caltech*1× training set). Results in italics indicate the use of matching training and test sets. The "pruned" variant improves performance for both detectors.

new annotations are indeed better aligned, just as *INRIA* annotations are better aligned than *Caltech*.

Decoupling classification from localisation errors. Next, we examine the impact of the new annotations on detection quality. We train *ACF* (Dollár *et al.*, 2014) and *RotatedFilters* models using different training sets and evaluate against both original and new annotations (i.e. MR^O_{-2}, MR^O_{-4} and MR^N_{-2}, MR^N_{-4}).

To help us decouple the effect of labelling errors from the effect of alignment errors, we generate a set of "pruned" annotations which addresses false positives and false negatives in the original set without improving alignment. To this end, we match new and original annotations with a criterion of IoU ≥ 0.5. Then we (i) mark as "ignore regions" any unmatched original annotations, and (ii) add new annotations absent in the original set. The resulting set of annotations can be viewed as a midpoint between original and new sets.

Tab. 6.5 shows results when trained with original, new, and pruned annotations using our *Caltech* training/validation split of $^5/_6 + ^1/_6$. As expected, models trained on original/new and tested on original/new perform better than training and testing on different annotations. Since the pruned annotations address labelling errors, these have a strong impact on the MR^O_{-2} which doesn't reflect alignment errors as strongly

Test proposals	Proposal	+AlexNet	+VGG	+bbox reg & NMS
ACF (Dollár *et al.*, 2014)	48.0%	28.5%	22.8%	20.8%
SquaresChnFtrs (Benenson *et al.*, 2014)	31.3%	21.2%	15.9%	14.7%
LDCF (Nam *et al.*, 2014)	23.7%	21.6%	16.0%	13.7%
RotatedFilters	17.2%	21.5%	17.8%	13.8%
Checkerboards (Zhang *et al.*, 2015b)	16.1%	21.0%	15.3%	11.1%
RotatedFilters-New10×	13.0%	17.2%	11.7%	10.0%

Table 6.6: Detection quality of CNNs with different proposal methods. Grey numbers indicate worse results than the input proposals. All numbers are reported in terms of MR_{-2}^{N} on the *Caltech* test set. The last column indicates bounding box regression followed by a second non-maximum suppression step after *VGG16* re-scoring.

as discussed above. We also observe in the "MR_{-2}^{N}" column that the stronger detector benefits more from better data.

Conclusion. Using high quality annotations for training improves the overall detection quality, thanks both to improved alignment and to reduced annotation errors.

6.5.2 CNNs for pedestrian detection

The results of Sec. 6.3.2 indicate that we can improve results by focusing on the classification subtask of detection. Chapter 5 and other recent work (Tian *et al.*, 2015b) have demonstrated competitive performance with convolutional neural networks (CNNs) for pedestrian detection. As these can help address classification errors, what is their behaviour w.r.t. localisation accuracy? To what extent is performance driven by the quality of the detection proposals?

AlexNet and VGG16. We consider two CNNs for detection: (i) *AlexNet* (used in Chapter 5), and (ii) the *VGG16* model used by Girshick (2015). Both are pre-trained on *ImageNet* and fine-tuned with *Caltech*10× (original annotations) using *SquaresChnFtrs* proposals. Both are instantiations of the *R-CNN* framework (Girshick *et al.*, 2014), albeit with slightly different training/test-time approaches (vanilla *R-CNN* versus *Fast R-CNN*). Nonetheless, we expect differences in the results to be dominated by the discriminative power of the respective CNNs. *VGG16* for example improves over *AlexNet* by 8pp (mAP) on the *PASCAL VOC* detection task (Girshick *et al.*, 2014).

Tab. 6.6 shows that as the quality of the detection proposals improves, *AlexNet* fails to provide consistent gains, eventually performing worse than the *ICF*-based proposal method. *VGG16* on the other hand almost consistently improves over the proposal method, but the gains shrink as the proposals improve.

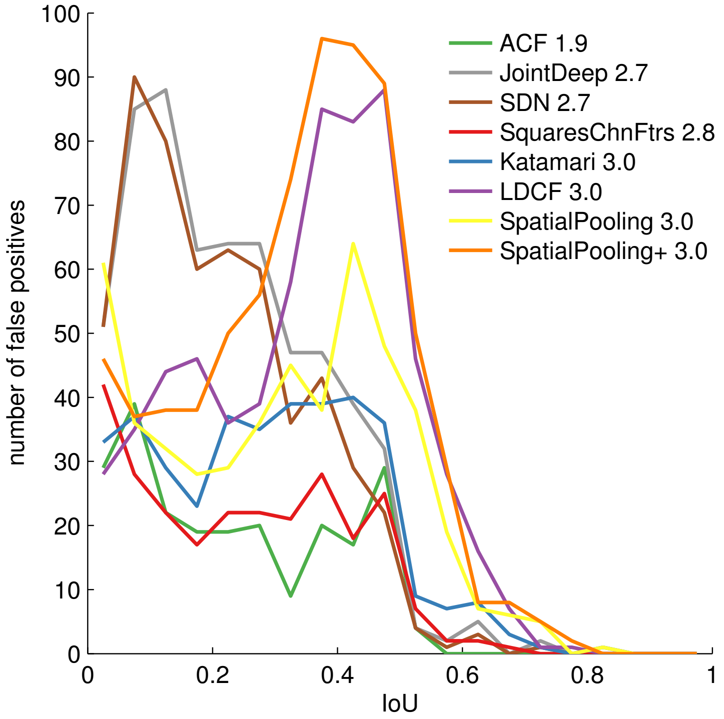


Figure 6.17: Distribution of overlap between false positives and ground truth, for different *ICF* detectors. The curves are histograms with coarse IoU bins (0 overlap case omitted). Number in the legend indicates the average number of proposals per image (after filtering to reach ∼3 proposals per image on average). Note that most detectors have many false positives near true detections.

By inspecting the resulting curves, we notice that both *AlexNet* and *VGG16* lower the scores of negative hypotheses but also generate a large number of high-scoring false positives. To get to the bottom of this, we look at the distribution of proposals. We find that *ICF* detectors are able to provide a set of proposals with high recall but also introduce many redundant detections that surround pedestrians (see Fig. 6.17). CNNs struggle to suppress the latter, as they produce more diffuse score maps. We hypothesise this is an intrinsic limitation of the *AlexNet* and *VGG16* architectures, due to the heavy reliance on subsampling operations during feature extraction. Obtaining "peakier" responses from a CNN will most likely require using rather different architectures, possibly ones that are designed for dense pixel-wise prediction tasks, such as semantic labelling or boundary detection.

Fortunately, we can compensate for the imprecise score maps by resorting to bounding box regression. We add such a regressor to *VGG16* and a second round of non-maximum suppression (NMS) separate from the one applied to the proposals. We use the usual IoU ≥ 0.5 merging criterion for the second NMS round. Neighbouring proposals that previously resulted in strong false positives are now combined into a single high-scoring detection. The last column of Tab. 6.6 demonstrates the resulting gains even over the best proposal method *RotatedFilters-New10*×. Evaluated against the original annotations,

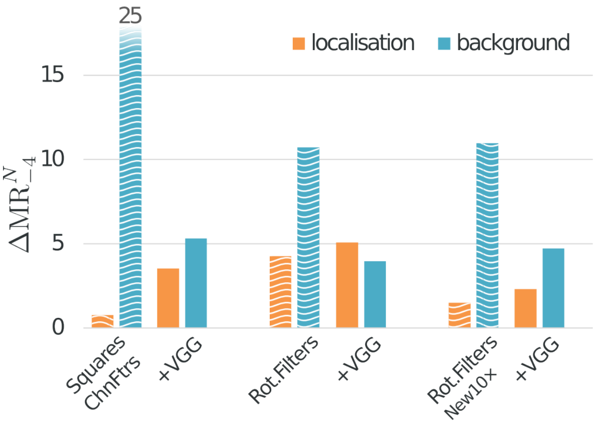


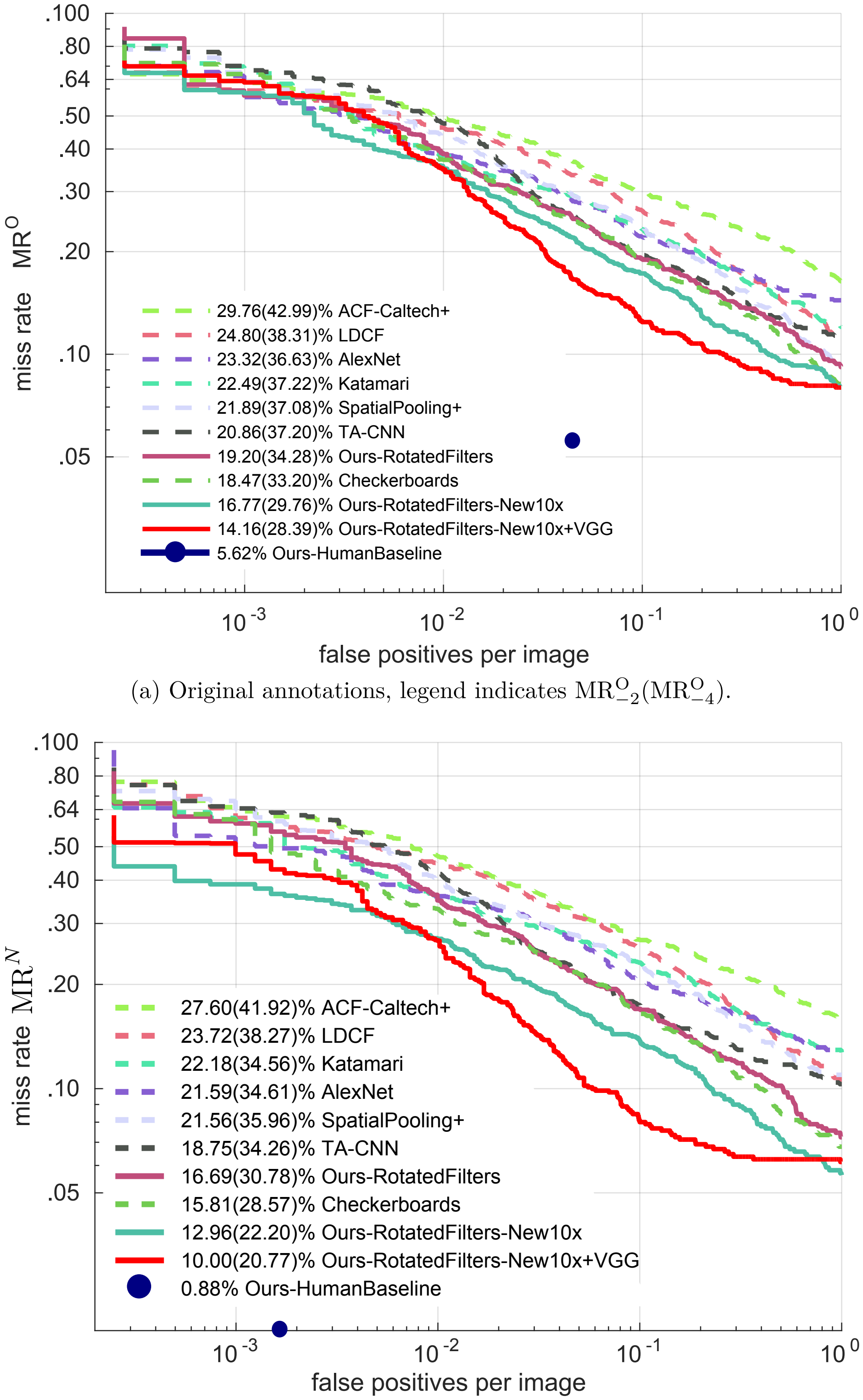
Figure 6.18: Oracle case analysis of proposals + CNNs (after a second round of NMS). The gain in miss rate is reported with ΔMR^{O}_{-4}. The CNN significantly improves background errors, while slightly increasing localisation errors.

RotatedFilters-New10×+VGG reaches 14.2% MR^{O}_{-2}, which improves over Chapter 5 and Tian *et al.* (2015b) as well as other state-of-the-art detectors (Fig. 6.19).

Fig. 6.18 repeats the oracle tests of Sec. 6.3.3 over our CNN results. We make comparisons for three CNN detectors and their corresponding *ICF* proposal methods, to observe how localisation and background errors change after *VGG16* re-scoring. One can see that for each proposal method, *VGG16* significantly cuts down on the background errors, while at the same time slightly increasing localisation errors. These comparisons verify that CNNs have strong discriminative ability against background objects, but on the other hand also demonstrate that CNNs fail to reduce the number of false positives close to ground truth objects.

Runtime comparisons. Our best performing detector *RotatedFilters-New10×+VGG* runs on a 640 × 480 image for ~3.5 seconds, including the *ICF* sliding window detection and *VGG16* re-scoring. Training *RotatedFilters* and fine-tuning *VGG16* each require 1~2 days. We compare the runtime versus performance for different detectors in Tab. 6.7. All detectors are tested on the same hardware: Intel Xeon E5-2680 2.70GHz CPU; and Tesla K40 GPU. Although *RotatedFilters-New10x+VGG* runs slower than previous *ICF* detectors, it reduces the errors by a large margin.

Conclusion. Although CNNs achieve strong results in image classification and general object detection, they seem to have limitations when it comes to producing well localised detection scores around small objects. Bounding box regression and NMS are key to addressing this limitation with current architectures. Despite this issue, classification remains the main source of errors, suggesting that there is still room for improvement for CNNs here as well.



(a) Original annotations, legend indicates $\mathrm{MR}^{\mathrm{O}}_{-2}(\mathrm{MR}^{\mathrm{O}}_{-4})$.

(b) New annotations, legend indicates $\mathrm{MR}^{\mathrm{N}}_{-2}(\mathrm{MR}^{\mathrm{N}}_{-4})$.

Figure 6.19: Performance of top detectors evaluated on original and new annotations.

	Runtime (seconds)			MR^N_{-2}
	CPU	GPU	Total	
ACF	0.1	/	0.1	27.6
Checkerboards	3.0	/	3.0	15.8
RotatedFilters-New10x	2.5	/	2.5	13.0
RotatedFilters-New10x+VGG	2.5	1.0	3.5	10.0

Table 6.7: Comparison of runtime versus performance for different detectors on the Caltech benchmark. Runtime is the average test time on one 640 × 480 image.

Detector aspect	MR^O_{-2} (MR^O_{-4})	MR^N_{-2} (MR^N_{-4})
Checkerboards	18.47 (33.20)	15.81 (28.57)
RotatedFilters	19.20 (34.28)	17.22 (31.65)
+ Alignment Sec. 6.5.1	16.97 (28.01)	14.54 (25.06)
+ New annotations Sec. 6.5.1	16.77 (29.76)	12.96 (22.20)
+ VGG Sec. 6.5.2	16.61 (34.79)	11.74 (28.37)
+ bbox reg & NMS	*14.16 (28.39)*	*10.00 (20.77)*

Table 6.8: Step by step improvements from previous best method *Checkerboards* to *RotatedFilters-New10x+VGG*.

6.6 Conclusion

In this chapter, we analysed the failures of top-performing detectors on the *Caltech* and *KITTI* datasets. With our human baseline, we have provided a lower bound on how much improvement there is to be expected on *Caltech*. There is a 10× gap in terms of errors still to be closed. To better measure the next steps in detection progress, we have provided new sanitised *Caltech* training and test set annotations.

Through a careful manual analysis, we identified different types of errors, which lead to specific suggestions on how to engineer better detectors (mentioned in Sec. 6.3.2; e.g. data augmentation for side-view persons, or extending the detector receptive field along the vertical axis).

We have partially addressed some of the issues by measuring the impact of better annotations on localisation accuracy, and by investigating the use of CNNs to improve the background to foreground discrimination. Our results indicate that significantly better alignment can be achieved with properly trained *ICF* detectors, and that, for pedestrian detection, CNNs struggle with localisation issues, which can be partially addressed via bounding box regression. Both on original and new annotations, the described detection approach reaches top performance, see progress in Tab. 6.8.

We hope the insights and data provided in this work will guide the path to closing the gap between machines and humans in the pedestrian detection task.

Part II

PIXELWISE LABELING

The Cityscapes Dataset for Semantic Urban Scene Understanding

VISUAL understanding of complex urban street scenes is an enabling factor for a wide range of applications. Object detection has benefited enormously from large-scale datasets, especially in the context of deep learning. For semantic urban scene understanding, there was a lack of datasets that adequately captured the complexity of real-world urban scenes.

To address this, we introduced *Cityscapes*, a benchmark suite and large-scale dataset to train and test approaches for pixel-level and instance-level semantic labelling. *Cityscapes* is comprised of a large, diverse set of stereo video sequences recorded in streets from 50 different cities. 5000 of these images have high quality pixel-level annotations; 20 000 additional images have coarse annotations to enable methods that leverage large volumes of weakly-labelled data. Crucially, our effort exceeded previous attempts in terms of dataset size, annotation richness, scene variability, and complexity. Our accompanying empirical study provides an in-depth analysis of the dataset characteristics, as well as a performance evaluation of several state-of-the-art approaches based on our benchmark. The dataset has since established itself as a go-to benchmark for pixel-level and instance-level semantic segmentation, but also has found use in other problem areas such as domain adaptation and generative modelling.

This work was published at CVPR (Cordts *et al.*, 2016). Marius Cordts was the lead author and Mohamed Omran contributed towards the defining and collecting the annotations, compiling dataset statistics, setting up the benchmark and the instance-level segmentation experiments.

7.1 Introduction

Visual scene understanding has moved from an elusive goal to a focus of much recent research in computer vision (Hoiem *et al.*, 2015). Semantic reasoning about the contents of a scene is thereby done on several levels of abstraction. Scene recognition aims to determine the overall scene category by putting emphasis on understanding its global properties, e.g. Zhou *et al.* (2014); Oliva and Torralba (2001). Scene labelling methods, on the other hand, seek to identify the individual constituent parts of a whole scene as well as their interrelations on a more local pixel- and instance-level, e.g. Long *et al.* (2015); Tighe *et al.* (2015). Specialized object-centric methods fall somewhere in

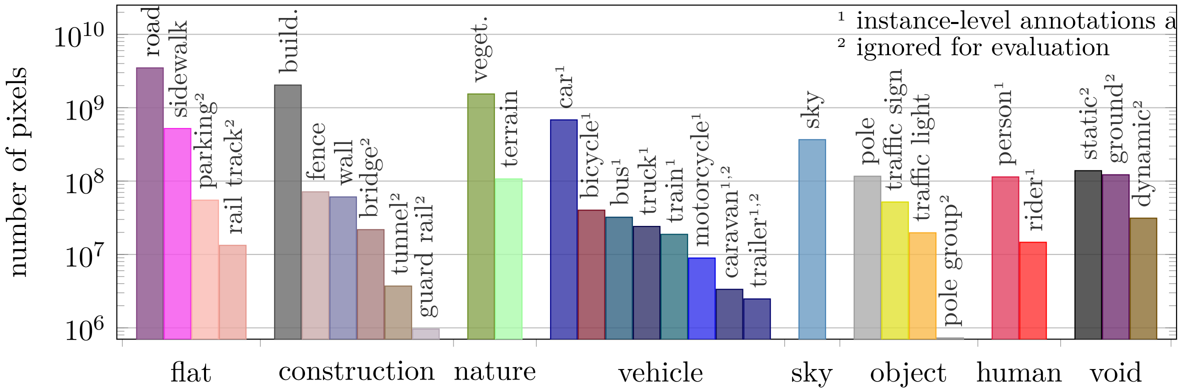


Figure 7.1: Number of finely annotated pixels (y-axis) per class and their associated categories (x-axis).

between by focusing on detecting a certain subset of (mostly dynamic) scene constituents, e.g. Felzenszwalb *et al.* (2010); Dollár *et al.* (2012b); Enzweiler and Gavrila (2009); Benenson *et al.* (2012). Despite significant advances, visual scene understanding remains challenging, particularly when taking human performance as a reference.

The resurrection of deep learning (LeCun *et al.*, 2015) has had a major impact on the current state-of-the-art in machine learning and computer vision. Many top-performing methods in a variety of applications are nowadays built around deep neural networks (Krizhevsky *et al.*, 2012; Sermanet *et al.*, 2014; Long *et al.*, 2015). A major contributing factor to their success is the availability of large-scale, publicly available datasets such as *ImageNet* (Russakovsky *et al.*, 2015a), *PASCAL VOC* (Everingham *et al.*, 2015), *PASCAL-Context* (Mottaghi *et al.*, 2014), and *Microsoft COCO (MSCOCO)* (Lin *et al.*, 2014) that allow deep neural networks to develop their full potential.

Despite the existing gap to human performance, scene understanding approaches have started to become essential components of advanced real-world systems. A particularly popular and challenging application involves self-driving cars, which make extreme demands on system performance and reliability. Consequently, significant research efforts have gone into new vision technologies for understanding complex traffic scenes and driving scenarios (Franke *et al.*, 2013; Furgale *et al.*, 2013; Geiger *et al.*, 2014; Scharwächter *et al.*, 2014; Ros *et al.*, 2015; Badrinarayanan *et al.*, 2017).

Also in this area, research progress can be heavily linked to the existence of datasets such as the *KITTI Vision Benchmark Suite* (Geiger *et al.*, 2013), *CamVid* (Brostow *et al.*, 2009), *Leuven* (Leibe *et al.*, 2007), and *Daimler Urban Segmentation* (Scharwächter *et al.*, 2013) datasets. These urban scene datasets are often much smaller than datasets addressing more general settings. Moreover, we argue that they do not fully capture the variability and complexity of real-world inner-city traffic scenes. Both shortcomings currently inhibit further progress in visual understanding of street scenes. To this end, we propose the *Cityscapes* benchmark suite and a corresponding dataset, specifically tailored for autonomous driving in an urban environment and involving a much wider range of highly complex inner-city street scenes that were recorded in 50 different cities.

Cityscapes significantly exceeds prior efforts in terms of size, annotation richness, and, more importantly, regarding scene complexity and variability. We go beyond pixel-level semantic labelling by also considering instance-level semantic labelling in both our annotations and evaluation metrics. To facilitate research on 3D scene understanding, we also provide depth information through stereo vision.

Concurrently with this work, Xie *et al.* (2016) announced a new semantic scene labelling dataset for suburban traffic scenes. It provides temporally consistent 3D semantic instance annotations with 2D annotations obtained through back-projection. We consider our efforts to be complementary given the differences in the way that semantic annotations are obtained, and in the type of scenes considered, i.e. suburban vs. inner-city traffic. To maximize synergies between both datasets, a common label definition that allows for cross-dataset evaluation has been mutually agreed upon and implemented.

7.2 Dataset

Designing a large-scale dataset requires a multitude of decisions, e.g. on the modalities of data recording, data preparation, and the annotation protocol. Our choices were guided by the ultimate goal of enabling significant progress in the field of semantic urban scene understanding.

7.2.1 Data specifications

Our data recording and annotation methodology was carefully designed to capture the high variability of outdoor street scenes. Several hundreds of thousands of frames were acquired from a moving vehicle during the span of several months, covering spring, summer, and fall in 50 cities, mostly in Germany but also in a couple of neighbouring

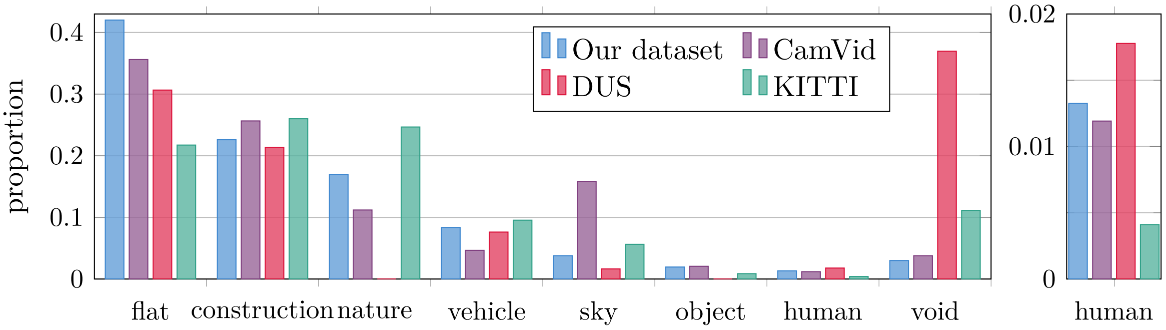


Figure 7.2: Proportion of annotated pixels (y-axis) per category (x-axis) for *Cityscapes*, *CamVid* (Brostow *et al.*, 2009), *DUS* (Scharwächter *et al.*, 2013), and *KITTI* (Geiger *et al.*, 2013).

countries. We deliberately did not record in adverse weather conditions, such as heavy rain or snow, as we believe such conditions require specialized techniques and datasets (Pfeiffer *et al.*, 2013).

Our camera system and post-processing reflect the current state-of-the-art in the automotive domain. Images were recorded with an automotive-grade 22 cm baseline stereo camera using 1/3 in CMOS 2 MP sensors (OnSemi AR0331) with rolling shutters at a frame-rate of 17 Hz. The sensors were mounted behind the windshield and yield high dynamic-range (HDR) images with 16 bits linear colour depth. Each 16-bit stereo image pair was subsequently debayered and rectified. We relied on the method of Kruger *et al.* (2004) for extrinsic and intrinsic calibration. To ensure calibration accuracy we re-calibrated on-site before each recording session.

For comparability and compatibility with existing datasets we also provide low dynamic-range (LDR) 8-bit RGB images that are obtained by applying a logarithmic compression curve. Such tone mappings are common in automotive vision, since they can be computed efficiently and independently for each pixel. To facilitate highest annotation quality, we applied a separate tone mapping to each image. The resulting images are less realistic, but visually more pleasing and proved easier to annotate. 5000 images were manually selected from 27 cities for dense pixel-level annotation, aiming for high diversity of foreground objects, background, and overall scene layout. The annotations (see Sec. 7.2.2) were done on the 20th frame of a 30-frame video snippet, which we provide in full to supply context information. For the remaining 23 cities, a single image every 20 s or 20 m driving distance (whatever comes first) was selected for coarse annotation, yielding 20 000 images in total.

In addition to the rectified 16-bit HDR and 8-bit LDR stereo image pairs and corresponding annotations, our dataset includes vehicle odometry obtained from in-vehicle sensors, outside temperature, and GPS tracks.

7.2.2 Classes and annotations

We provide coarse and fine annotations at the pixel level including instance-level labels for humans and vehicles.

Our 5000 fine pixel-level annotations consist of layered polygons à la *LabelMe* (Russell *et al.*, 2008). These were collected in-house to guarantee the highest levels of quality. Annotation and quality control required more than 1.5 h on average for a single image. Annotators were asked to label the image from back to front such that no object boundary was marked more than once. Each annotation thus implicitly provides a depth ordering of the objects in the scene. Given our labelling scheme, annotations can be easily extended to cover additional or more fine-grained classes.

For our 20 000 coarse pixel-level annotations, accuracy on object boundaries was traded off against annotation speed. We aimed to correctly annotate as many pixels as possible within a 7 min window per image. This was achieved by labelling coarse polygons under the sole constraint that each polygon must only include pixels belonging to a single object class.

We assessed the quality of our labelling with two experiments. In the first, 30 images were finely annotated twice by different annotators. A comparison of the two sets of annotations showed that 96 % of all pixels were assigned to the same label. Our annotators were instructed to choose a *void* label if unsure of the object class. We thus repeated the comparison without pixels covered by at least one *void* polygon. This yielded a label agreement of 98 %. For the second experiment, we coarsely reannotated all images with fine annotations, with the purpose of enabling research on densifying coarse labels. Comparing coarse and fine annotations showed that 97 % of all coarsely labelled pixels were assigned the same class as in the fine annotations.

We defined 30 visual classes for annotation, which are grouped into eight categories: flat, construction, nature, vehicle, sky, object, human, and void. Classes were selected based on their frequency, relevance from an application standpoint, practical considerations regarding the annotation effort, as well as the desire to facilitate compatibility with existing datasets, e.g. Geiger *et al.* (2013); Brostow *et al.* (2009); Xie *et al.* (2016). Classes that ended up being rarely annotated were subsequently excluded from our benchmark, leaving 19 classes for evaluation, see Fig. 7.1 for details. Our annotation tool is also publicly available.[8]

7.2.3 Dataset splits

We split our densely annotated images into separate training, validation, and test sets. The coarsely annotated images are solely meant to serve as additional training data. We chose not to split the data randomly, but rather in a way that ensures each split is representative of the variability of different street scene scenarios. The underlying split criteria were chosen to ensure a roughly balanced distribution of geographic location and population size of the individual cities, as well as of the time of year during which recordings took place. Specifically, each of the three split sets is comprised of data recorded with the following properties in equal shares: (i) in large, medium, and small cities; (ii) in the geographic west, centre, and east; (iii) in the geographic north, centre, and south; (iv) at the beginning, middle, and end of the year. Note that the data is split at the city level, i.e. the images recorded in a single city are completely contained within a single split. Following this scheme, we arrive at a unique split consisting of 2975 training and 500 validation images with publicly available annotations, as well as 1525 test images with annotations withheld for benchmarking purposes.

In order to assess how uniform (representative) the splits are regarding the four split characteristics, we trained a fully convolutional network (Long *et al.*, 2015) on the 500 images in our validation set. This model was then evaluated on the whole test set, as well as on the eight subsets thereof that reflect the extreme values of the four characteristics.

[8] http://github.com/mcordts/cityscapesScripts

	#pixels [10^9]	annot. density [%]
Ours (fine)	9.43	**97.1**
Ours (coarse)	**26.0**	67.5
CamVid	0.62	96.2
DUS	0.14	63.0
KITTI	0.23	88.9

Table 7.1: Absolute number and density of annotated pixels for *Cityscapes*, *DUS*, *KITTI*, and *CamVid* (upscaled to 1280 × 720 pixels to maintain the original aspect ratio).

	#humans [10^3]	#vehicles [10^3]	#h/image	#v/image
Ours (fine)	24.4	**41.0**	**7.0**	**11.8**
KITTI	6.1	30.3	0.8	4.1
Caltech	**192**[1]	-	1.5	-

Table 7.2: Absolute and average number of instances (humans and vehicles) for *Cityscapes*, *KITTI*, and *Caltech* ([1] via interpolation) on the respective training and validation datasets.

With the exception of the time of year, the performance is very homogeneous, varying less than 1.5 % points (often much less). Interestingly, the performance on the *end of the year* subset is 3.8 % points better than on the whole test set. We hypothesise that this is due to softer lighting conditions in the frequently cloudy fall. To verify this hypothesis, we additionally tested on images taken in low- or high-temperature conditions, finding a 4.5 % point increase in low temperatures (cloudy) and a 0.9 % point decrease in warm (sunny) weather. Moreover, specifically training for either condition leads to an improvement on the respective test set, but not on the balanced set. These findings support our hypothesis and underline the importance of a dataset covering a wide range of conditions encountered in the real world in a balanced way.

7.2.4 Statistical analysis

We compare *Cityscapes* to other datasets in terms of (i) annotation volume and density, (ii) the distribution of visual classes, and (iii) scene complexity. Regarding the first two aspects, we compare *Cityscapes* to other datasets with semantic pixel-wise annotations, i.e. *CamVid* (Brostow *et al.*, 2009), *DUS* (Scharwächter *et al.*, 2014), and *KITTI* (Geiger *et al.*, 2013). Note that there are many other datasets with dense semantic annotations, e.g. Ardeshir *et al.* (2015); Song *et al.* (2015); Sengupta *et al.* (2012); Riemenschneider *et al.* (2014); Tighe and Lazebnik (2013). However, we restrict this part of the analysis to those with a focus on autonomous driving.

CamVid consists of ten minutes of video footage with pixel-wise annotations for over 700 frames. *DUS* consists of a video sequence of 5000 images from which 500 have been annotated. *KITTI* addresses several different tasks including semantic labelling and object detection. As no official pixel-wise annotations exist for *KITTI*, several independent groups have annotated approximately 700 frames (Xu *et al.*, 2013; Sengupta *et al.*, 2013; He and Upcroft, 2013; Ladicky *et al.*, 2014; Kundu *et al.*, 2014; Ros *et al.*, 2015; Güney and Geiger, 2015; Zhang *et al.*, 2015a). We map those labels to our high-level categories and analyse this consolidated set. In comparison, *Cityscapes* provides significantly more annotated images, i.e. 5000 fine and 20 000 coarse annotations. Moreover, the annotation quality and richness is notably better. As *Cityscapes* provides recordings from 50 different cities, it also covers a significantly larger area than previous datasets that contain images from a single city only, e.g. Cambridge (*CamVid*), Heidelberg (*DUS*), and Karlsruhe (*KITTI*). In terms of absolute and relative numbers of semantically annotated pixels (training, validation, and test data), *Cityscapes* compares favourably to *CamVid*, *DUS*, and *KITTI* with up to two orders of magnitude more annotated pixels, cf. Tab. 7.1. The majority of all annotated pixels in *Cityscapes* belong to the coarse annotations, providing many individual (but correlated) training samples, but missing information close to object boundaries.

Figs. 7.1 and 7.2 compare the distribution of annotations across individual classes and their associated higher-level categories. Notable differences stem from the inherently different configurations of the datasets. *Cityscapes* involves dense inner-city traffic with wide roads and large intersections, whereas *KITTI* is composed of less busy suburban traffic scenes. As a result, *KITTI* exhibits significantly fewer “flat” ground structures, fewer “humans”, and more “nature”. In terms of overall composition, *DUS* and *CamVid* seem more aligned with *Cityscapes*. Exceptions are an abundance of “sky” pixels in *CamVid* due to cameras with a comparably large vertical field-of-view and the absence of certain categories in *DUS*, i.e. “nature” and “object“.

Finally, we assess scene complexity, where density and scale of traffic participants (humans and vehicles) serve as proxy measures. Out of the previously discussed datasets, only *Cityscapes* and *KITTI* provide instance-level annotations for humans and vehicles. We additionally compare to the *Caltech Pedestrian Dataset* (Dollár *et al.*, 2012b), which only contains annotations for humans, but none for vehicles. Furthermore, *KITTI* and *Caltech* only provide instance-level annotations in terms of axis-aligned bounding boxes. We use the respective training and validation splits for our analysis, since test set annotations are not publicly available for all datasets. In absolute terms, *Cityscapes* contains significantly more object instance annotations than *KITTI*, see Tab. 7.2. Being a specialised benchmark, *Caltech* provides the most annotations for humans by a margin. The major share of those labels was obtained, however, by interpolation between a sparse set of manual annotations resulting in significantly degraded label quality. The relative statistics emphasize the much higher complexity of *Cityscapes*, as the average numbers of object instances per image notably exceed those of *KITTI* and *Caltech*. We extend our analysis to *MSCOCO* (Lin *et al.*, 2014) and *PASCAL VOC* (Everingham *et al.*, 2015), which also contain street scenes while not being restricted to them. We analyse the frequency of scenes with a certain number of traffic participants, see Fig. 7.3. We find that our dataset covers a greater range of scene complexity and has a higher

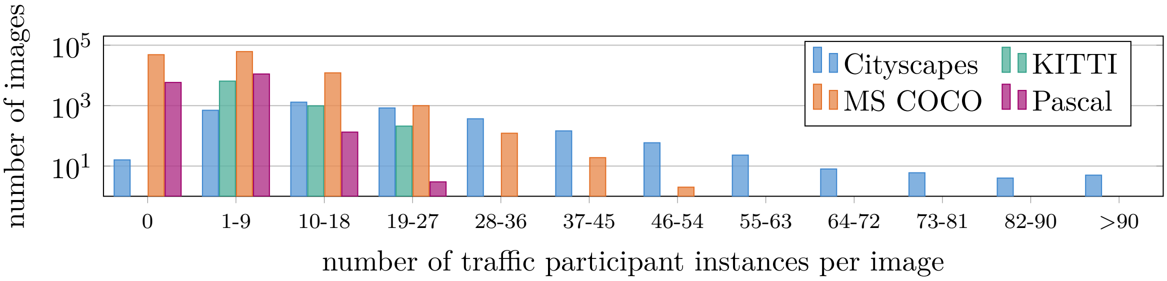


Figure 7.3: Dataset statistics regarding scene complexity. Only MS COCO and Cityscapes provide instance segmentation masks.

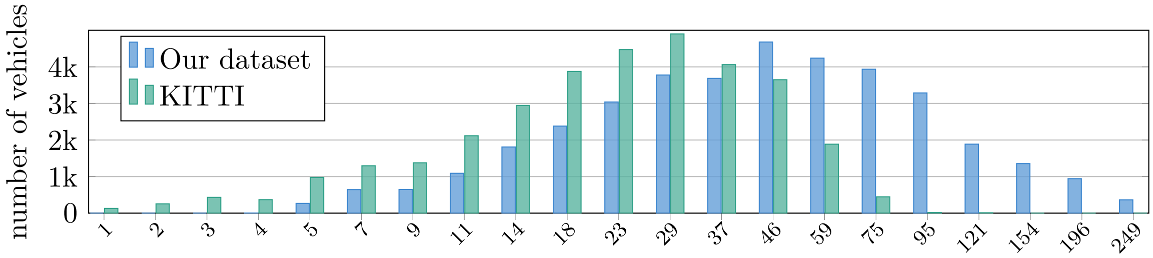


Figure 7.4: Histogram of object distances in meters for class *vehicle*.

proportion of highly complex scenes compared to previous datasets. Using stereo data, we analyse the distribution of vehicle distances to the camera. From Fig. 7.4 we observe that in comparison to *KITTI*, *Cityscapes* covers a larger distance range. We attribute this to both our higher-resolution imagery and the careful annotation procedure. As a consequence, algorithms need to account for a larger range of scales and object sizes to score well in our benchmark.

7.3 Semantic Labelling

The first *Cityscapes* task involves predicting a per-pixel semantic labelling of the image without considering higher-level object instance or boundary information.

7.3.1 Tasks and metrics

To assess labelling performance, we rely on one standard and one novel metric. The first is the standard Jaccard Index, commonly known as the *PASCAL VOC* intersection-over-union metric $\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}}$ (Everingham *et al.*, 2015), where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels, respectively, determined over the whole test set. Owing to the two levels of semantic granularity, i.e. classes and categories, we report two separate mean performance scores: $\mathrm{IoU}_{\mathrm{category}}$ and $\mathrm{IoU}_{\mathrm{class}}$. In either case, pixels labelled as void do not contribute to the score.

The global IoU measure is biased toward object instances that cover a large image area. In street scenes with their strong scale variation this can be problematic. Specifically for traffic participants, which are the key classes in our scenario, we aim to evaluate how well the individual instances in the scene are represented in the labelling. To address this, we additionally evaluate the semantic labelling using an instance-level intersection-over-union metric $\mathrm{iIoU} = \frac{\mathrm{iTP}}{\mathrm{iTP}+\mathrm{FP}+\mathrm{iFN}}$. Here, iTP, and iFN denote weighted counts of true positive and false negative pixels, respectively. In contrast to the standard IoU measure, the contribution of each pixel is weighted by the ratio of a class's average instance size to the size of the respective ground truth instance. As before, FP is the number of false positive pixels. It is important to note here that unlike the instance-level task in Sec. 7.4, we assume that the methods only yield a standard per-pixel semantic class labelling as output. Therefore, the false positive pixels are not associated with any instance and thus do not require normalization. The final scores, $\mathrm{iIoU}_{\mathrm{category}}$ and $\mathrm{iIoU}_{\mathrm{class}}$, are obtained as the respective means for the two levels of semantic granularity, while only classes with instance annotations are included.

7.3.2 Control experiments

We conduct several control experiments to put our baseline results below into perspective. First, we count the relative frequency of every class label at each pixel location of the fine (coarse) training annotations. Using the most frequent label at each pixel as a constant prediction irrespective of the test image (called *static fine* (SF) and *static coarse* (SC) respectively) results in roughly 10 % $\mathrm{IoU}_{\mathrm{class}}$, as shown in Tab. 7.3. These low scores emphasize the high diversity of our data. SC and SF having similar performance indicates the value of our additional coarse annotations. Even if the ground truth (GT) segments are re-classified using the most frequent training label (SF or SC) within each segment mask, the performance does not notably increase.

Secondly, we re-classify each ground truth segment using *FCN-8s* (Long *et al.*, 2015), cf. Sec. 7.3.4. We compute the average scores within each segment and assign the maximizing label. The performance is significantly better than the static predictors but still far from 100 %. We conclude that it is necessary to optimise both classification and segmentation quality at the same time.

Thirdly, we evaluate the performance of subsampled ground truth annotations as predictors. Subsampling was done by majority voting of neighbouring pixels, followed by resampling back to full resolution. This yields an upper bound on the performance at a fixed output resolution and is particularly relevant for deep learning approaches that often apply downscaling due to constraints on time, memory, or the network architecture itself. Downsampling factors 2 and 4 correspond to the most common setting of our 3$^{\mathrm{rd}}$-party baselines (Sec. 7.3.4). Note that while subsampling by a factor of 2 hardly affects the IoU score, it clearly decreases the iIoU score given its comparatively large impact on small, but nevertheless important objects. This underlines the importance of the separate instance-normalised evaluation. The downsampling factors of 8, 16, and 32 are motivated by the corresponding strides of the FCN model. The performance of a GT downsampling by a factor of 64 is comparable to the current state of the art, while downsampling by a factor of 128 is the smallest (power of 2) downsampling for which all images have a distinct labelling.

Lastly, we employ 128-times subsampled annotations and retrieve the nearest training annotation in terms of the Hamming distance. The full resolution version of this training annotation is then used as prediction, resulting in 21 % $\mathrm{IoU}_{\mathrm{class}}$. While outperforming the static predictions, the poor result demonstrates the high variability of our dataset and its demand for approaches that generalise well.

7.3.3 State of the art

Drawing on the success of deep learning algorithms, a number of semantic labelling approaches have shown very promising results and significantly advanced the state of the art. These new approaches take enormous advantage from recently introduced large-scale datasets, e.g. *PASCAL-Context* (Mottaghi *et al.*, 2014) and *Microsoft COCO (MSCOCO)* (Lin *et al.*, 2014). *Cityscapes* aims to complement these, particularly in the context of understanding complex urban scenarios, in order to enable further research in this area.

The popular work of Long *et al.* (2015) showed how a top-performing Convolutional Neural Network (CNN) for image classification can be successfully adapted for the task of semantic labelling by the careful use of upsampling layers. Similarly, Yu and Koltun (2016) adapt a classification CNN by introducing dilated convolutions that avoid a loss of resolution that results from sub-sampling pooling layers.

Several other methods propose to combine the strengths of CNNs and Conditional Random Fields (CRFs) (Chen *et al.*, 2015a; Zheng *et al.*, 2015; Schwing and Urtasun, 2015; Liu *et al.*, 2015b; Lin *et al.*, 2016).

Other work takes advantage of deep learning for explicitly integrating global scene context in the prediction of pixel-wise semantic labels, in particular through CNNs (Liu *et al.*, 2015a; Badrinarayanan *et al.*, 2017; Mostajabi *et al.*, 2015; Sharma *et al.*, 2015) or Recurrent Neural Networks (RNNs) (Pinheiro and Collobert, 2014; Byeon *et al.*, 2015). Last but not least, several recent studies have explored different forms of weak supervision, such as bounding boxes or image-level labels, for training CNNs for pixel-level semantic labelling (Papandreou *et al.*, 2015; Dai *et al.*, 2015; Pinheiro and Collobert, 2015; Xu *et al.*, 2015; Pathak *et al.*, 2015b,a; Bearman *et al.*, 2016; Wei *et al.*, 2017). We hope our coarse annotations can further advance this area.

Average over	Classes		Categories	
Metric [%]	IoU	iIoU	IoU	iIoU
static fine (SF)	10.1	4.7	26.3	19.9
static coarse (SC)	10.3	5.0	27.5	21.7
GT segmentation with SF	10.1	6.3	26.5	25.0
GT segmentation with SC	10.9	6.3	29.6	27.0
GT segmentation with Long *et al.* (2015)	79.4	52.6	93.3	80.9
GT subsampled by 2	97.2	92.6	97.6	93.3
GT subsampled by 4	95.2	90.4	96.0	91.2
GT subsampled by 8	90.7	82.8	92.1	83.9
GT subsampled by 16	84.6	70.8	87.4	72.9
GT subsampled by 32	75.4	53.7	80.2	58.1
GT subsampled by 64	63.8	35.1	71.0	39.6
GT subsampled by 128	50.6	21.1	60.6	29.9
nearest training neighbour	21.3	5.9	39.7	18.6

Table 7.3: Quantitative results of control experiments for semantic labelling using the metrics presented in Sec. 7.3.1.

7.3.4 Baselines

Our own baseline experiments (Tab. 7.4, top) rely on fully convolutional networks (FCNs), as they are central to most state-of-the-art methods (Long *et al.*, 2015; Schwing and Urtasun, 2015; Chen *et al.*, 2015a; Lin *et al.*, 2016; Zheng *et al.*, 2015). We adopted *VGG16* Simonyan and Zisserman (2015) and utilise the *PASCAL-Context* setup of Long *et al.* (2015) with a modified learning rate to match our image resolution under an unnormalised loss. According to the notation in Long *et al.* (2015), we denote the different models as *FCN-32s*, *FCN-16s*, and *FCN-8s*, where the numbers are the stride of the finest heatmap. Since *VGG16* training on 2 MP images exceeds even the largest GPU memory available, we split each image into two halves with sufficiently large overlap. Additionally, we trained a model on images downscaled by a factor of 2. We first train on our training set (*train*) until the performance on our validation set (*val*) saturates, and then retrain on *train+val* with the same number of epochs.

To obtain further baseline results, we asked selected groups that have proposed state-of-the-art semantic labelling approaches to optimise their methods on our dataset and evaluated their predictions on our test set. The resulting scores are given in Tab. 7.4 (bottom) and qualitative examples of three selected methods are shown in Fig. 7.5. Interestingly enough, the performance ranking in terms of the main $\mathrm{IoU}_{\mathrm{class}}$ score on Cityscapes is highly different from *PASCAL VOC* (Everingham *et al.*, 2015). While *DPN* is the 2nd best method on *PASCAL VOC*, it is only the 6th best on *Cityscapes*. *FCN-8s* is last on *PASCAL*, but 3rd best on *Cityscapes*. *Adelaide-CNN-CRF* performs consistently well on both datasets with rank 1 on *PASCAL* and 2 on *Cityscapes*.

	train	val	coarse	sub	Classes		Categories	
					IoU	iIoU	IoU	iIoU
FCN-32s	✓	✓			61.3	38.2	82.2	65.4
FCN-16s	✓	✓			64.3	41.1	84.5	69.2
FCN-8s	✓	✓			65.3	41.7	85.7	70.1
FCN-8s	✓	✓		2	61.9	33.6	81.6	60.9
FCN-8s		✓			58.3	37.4	83.4	67.2
FCN-8s			✓		58.0	31.8	78.2	58.4
SegNet-extended (Badrinarayanan *et al.*, 2017)	✓			4	56.1	34.2	79.8	66.4
SegNet-basic (Badrinarayanan *et al.*, 2017)	✓			4	57.0	32.0	79.1	61.9
DPN (Liu *et al.*, 2015b)	✓	✓	✓	3	59.1	28.1	79.5	57.9
CRFasRNN (Zheng *et al.*, 2015)	✓			2	62.5	34.4	82.7	66.0
DeepLab-CRF (Chen *et al.*, 2015a)	✓	✓		2	63.1	34.5	81.2	58.7
DeepLab-CRF-weaksup (Papandreou *et al.*, 2015)	✓	✓	✓	2	64.8	34.9	81.3	58.7
Adelaide-CNN-CRF (Lin *et al.*, 2016)	✓				66.4	**46.7**	82.8	67.4
Dilated10 (Yu and Koltun, 2016)	✓				**67.1**	42.0	**86.5**	**71.1**

Table 7.4: Quantitative results of baselines for semantic labelling using the metrics presented in Sec. 7.3.1. The first block lists results from our own experiments, the second from those provided by 3$^{\text{rd}}$ parties. All numbers are given in percent and we indicate the used training data for each method, i.e. train fine, val fine, coarse extra as well as a potential downscaling factor (sub) of the input image.

From examining these results, we draw several conclusions: (1) The amount of downscaling applied during training and testing has a strong and consistent negative influence on performance (cf. *FCN-8s* vs. *FCN-8s* at half resolution, as well as the 2$^{\text{nd}}$ half of the table). The ranking according to IoU$_{\text{class}}$ is strictly consistent with the degree of downscaling. We attribute this to the large scale variation present in our dataset, cf. Fig. 7.4. This observation clearly indicates the demand for additional research in the direction of memory and computationally efficient CNNs when facing such a large-scale dataset with high-resolution images. (2) Our novel iIoU metric treats instances of any size equally and is therefore more sensitive to errors in predicting small objects compared to the IoU. Methods that leverage a CRF (*CRFasRNN*, *DPN*, *DeepLab-CRF*, *DeepLab-CRF-weaksup*) for regularisation tend to over smooth small objects, cf. Fig. 7.5, hence show a larger drop from IoU to iIoU than *SegNet* or *FCN-8s*. *Adelaide-CNN-CRF* is the only exception; its specific FCN-derived pairwise terms apparently allow for a more selective regularisation. (3) When considering IoU$_{\text{category}}$, *Dilated10* and *FCN-8s* perform particularly well, indicating that these approaches produce comparatively many confusions between the classes within the same category, cf. the buses in Fig. 7.5 (top). (4) Training *FCN-8s* with 500 densely annotated images (750 h of annotation) yields comparable IoU performance to a model trained on 20 000 weakly annotated images (1300 h annot.), cf. rows 5 & 6 in Tab. 7.4. However, in both cases the performance is significantly lower than *FCN-8s* trained on all 3475 densely annotated images. Many fine labels are thus important for training standard methods as well as for testing, but the performance only using coarse annotations does not collapse and presents a viable option. (5) Since the coarse annotations do not include small or distant instances, their iIoU performance is worse. (6) Coarse labels can complement the dense labels if applying appropriate methods as evidenced by *DeepLab-CRF-weaksup* outperforming *DeepLab-CRF*, which it extends by exploiting both dense and weak annotations (e.g. bounding boxes). Our dataset will hopefully stimulate research on exploiting the coarse labels further, especially given the interest in this area, e.g. (Oquab *et al.*, 2015; Hattori *et al.*, 2015; Misra *et al.*, 2015).

Overall, we believe that the unique characteristics of our dataset (e.g. scale variation, amount of small objects, focus on urban street scenes) allow for more such novel insights.

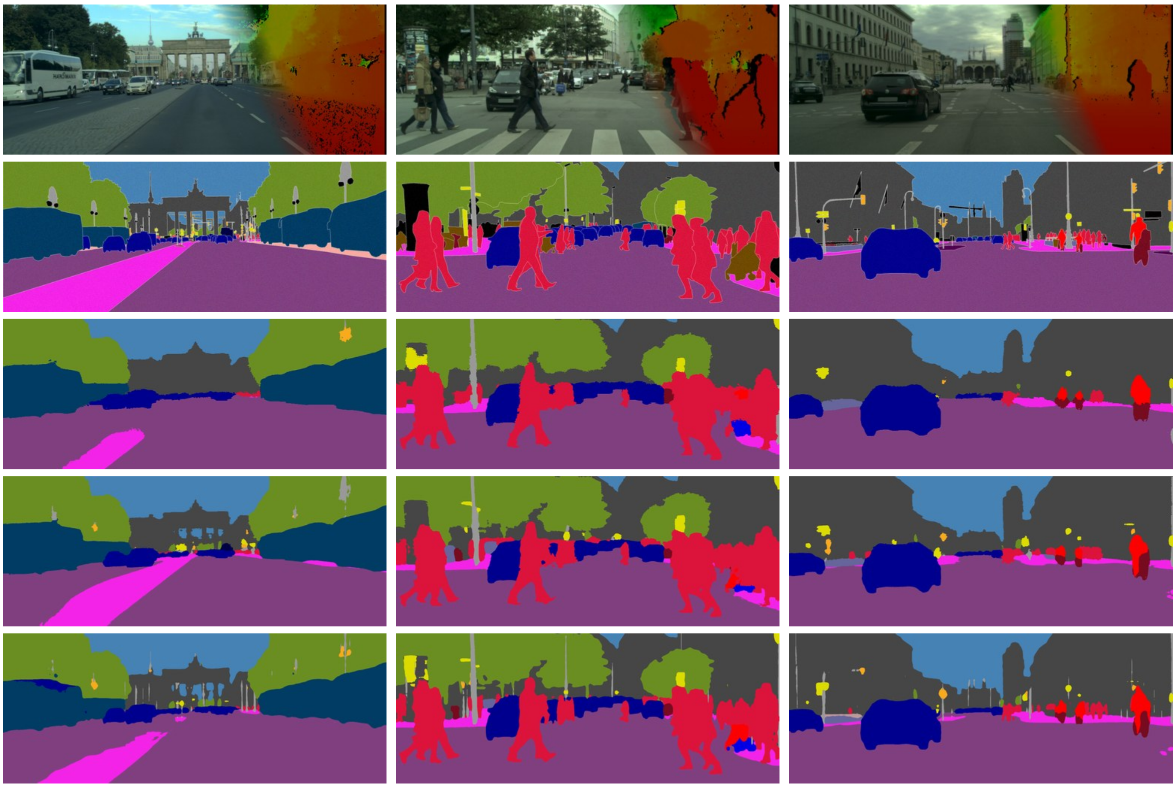


Figure 7.5: Qualitative examples of selected baselines. From top to bottom: (i) image with partially overlayed stereo depth maps, (ii) ground truth annotation, (iii) *DeepLab-CRF-weaksup* (Papandreou *et al.*, 2015), (iv) *Adelaide-CNN-CRF* (Lin *et al.*, 2016), and (v) *Dilated10* (Yu and Koltun, 2016). The colour coding of the semantic classes matches Fig. 7.1.

7.3.5 Cross-dataset evaluation

In order to show the compatibility and complementarity of *Cityscapes* regarding related datasets, we applied an FCN model trained on our data to *CamVid* Brostow *et al.* (2009) and two subsets of *KITTI* (Ros *et al.*, 2015; Sengupta *et al.*, 2013). We use

Dataset	Best reported result	Our result
CamVid (Brostow *et al.*, 2009)	62.9 (SegNet)	72.6
KITTI (Ros *et al.*, 2015)	61.6 (SegNet)	70.9
KITTI (Sengupta *et al.*, 2013)	82.2 (DenseSemFusion)	81.2

Table 7.5: Quantitative results (avg. recall in percent) of our half-resolution *FCN-8s* model trained on *Cityscapes* images and tested on *CamVid* and two pixel-wise labelled subsets of *KITTI*. We compare against the results of *SegNet* (Badrinarayanan *et al.*, 2017) and *DenseSemFusion* (Vineet *et al.*, 2015)

the half-resolution model (cf. 4th row in Tab. 7.4) to better match the target datasets, but we do not apply any specific training or fine-tuning. In all cases, we follow the evaluation of the respective dataset to be able to compare to previously reported results (Badrinarayanan *et al.*, 2017; Vineet *et al.*, 2015). The obtained results in Tab. 7.5 show that our large-scale dataset enables us to train models that are on a par with or even outperforming methods that are specifically trained on another benchmark and specialised for its test data. Further, our analysis shows that our new dataset integrates well with existing ones and allows for cross-dataset research.

7.4 Instance-Level Semantic Labelling

The pixel-level task, cf. Sec. 7.3, does not aim to segment individual object instances. In contrast, in the instance-level semantic labelling task, we focus on simultaneously detecting objects and segmenting them. This is an extension to both traditional object detection, since per-instance segments must be provided, and semantic labelling, since each instance is treated as a separate label.

7.4.1 Tasks and metrics

For instance-level semantic labelling, algorithms are required to deliver a set of detections of traffic participants in the scene, each associated with a confidence score and a per-instance segmentation mask. To assess instance-level performance, we compute the average precision on the region level (AP^r) (Hariharan *et al.*, 2014a) for each class and average it across a range of overlap thresholds to avoid a bias towards a specific value. Specifically, we follow Lin *et al.* (2014) and use 10 different overlaps ranging from 0.5 to 0.95 in steps of 0.05. The overlap is computed at the region level, making it equivalent to the IoU of a single instance. We penalise multiple predictions of the same ground truth instance as false positives. To obtain a single, easy to compare compound score, we report the mean average precision (mAP^r), obtained by also averaging over the class label set. As minor scores, we add $\mathrm{mAP}^r_{50\%}$ for an overlap value of 50 %, as well as $\mathrm{mAP}^r_{100\mathrm{m}}$ and $\mathrm{mAP}^r_{50\mathrm{m}}$ where the evaluation is restricted to objects within 100 m and 50 m distance, respectively.

7.4.2 State of the art

As detection results have matured (70 % AP on *PASCAL* (Everingham *et al.*, 2015; Ren *et al.*, 2015)), the last years have seen a rising interest in more difficult settings. Detections with pixel-level segments rather than traditional bounding boxes provide a richer output and allow (in principle) for better occlusion handling. We group existing methods into three categories.

The first encompasses **segmentation, then detection** and most prominently the *R-CNN* detection framework (Girshick *et al.*, 2014), relying on object proposals for generating detections. Many of the commonly used bounding box proposal methods (Hosang *et al.*, 2016; Pont-Tuset and Gool, 2015) first generate a set of overlapping segments, e.g. *Selective Search* (Uijlings *et al.*, 2013) or *MCG* (Arbeláez *et al.*, 2014). In *R-CNN*, bounding boxes of each segment are then scored using a CNN-based classifier, while each segment is treated independently.

The second category encompasses **detection, then segmentation**, where bounding-box detections are refined to instance specific segmentations. Either CNNs (Hariharan *et al.*, 2014a, 2015) or non-parametric methods (Chen *et al.*, 2015b) are typically used, however, in both cases without coupling between individual predictions.

Third, simultaneous **detection and segmentation** is significantly more delicate. Earlier methods relied on Hough voting (Leibe *et al.*, 2008; Riemenschneider *et al.*, 2012). More recent works formulate a joint inference problem on pixel and instance level using CRFs (Maire *et al.*, 2011; Yao *et al.*, 2012; He and Gould, 2014; Dai *et al.*, 2015; Tighe *et al.*, 2015; Zhang *et al.*, 2015c). Differences lie in the generation of proposals (exemplars, average class shape, direct regression), the cues considered (pixel-level labelling, depth ordering), and the inference method (probabilistic, heuristics).

7.4.3 Lower bounds, oracles & baselines

In Tab. 7.6, we provide lower-bounds that any sensible method should improve upon, as well as oracle-case results (i.e. using the test time ground truth). For our experiments, we rely on publicly available implementations. We train a *Fast R-CNN* detector (Girshick, 2015) on our training data in order to score *MCG* object proposals (Arbeláez *et al.*, 2014). Then, we use either its output bounding boxes as (rectangular) segmentations, the associated region proposal, or its convex hull as a per-instance segmentation. The best main score mAP^r is 4.6 %, is obtained with convex hull proposals, and becomes larger when restricting the evaluation to 50 % overlap or close instances. We contribute these rather low scores to our challenging dataset, biased towards busy and cluttered scenes, where many, often highly occluded, objects occur at various scales, cf. Sec. 7.2. Further, the *MCG* bottom-up proposals seem to be unsuited for such street scenes and cause extremely low scores when requiring large overlaps.

Proposals	Classif.	mAP^r	$\text{mAP}^r_{50\%}$	$\text{mAP}^r_{100\text{m}}$	$\text{mAP}^r_{50\text{m}}$
MCG regions	Fast R-CNN	2.6	9.0	4.4	5.5
MCG bboxes	Fast R-CNN	3.8	11.3	6.5	8.9
MCG hulls	Fast R-CNN	**4.6**	**12.9**	**7.7**	**10.3**
GT bboxes	Fast R-CNN	8.2	23.7	12.6	15.2
GT regions	Fast R-CNN	41.3	41.3	58.1	64.9
MCG regions	GT	10.5	27.0	16.0	18.7
MCG bboxes	GT	9.9	25.8	15.3	18.9
MCG hulls	GT	11.6	29.1	17.7	21.4

Table 7.6: Baseline results on instance-level semantic labelling task using the metrics described in Sec. 7.4. All numbers in %.

We confirm this interpretation with oracle experiments, where we replace the proposals at test-time with ground truth segments or replace the *Fast R-CNN* classifier with an oracle. In doing so, the task of object localization is decoupled from the classification task. The results in Tab. 7.6 show that when bound to *MCG* proposals, the oracle classifier is only slightly better than *Fast R-CNN*. On the other hand, when the proposals are perfect, *Fast R-CNN* achieves decent results. Overall, these observations unveil that the instance-level performance of our baseline is bound by the region proposals.

7.5 Conclusion

In this chapter, we presented *Cityscapes*, a comprehensive benchmark suite that has been carefully designed to spark progress in semantic urban scene understanding by: (i) creating the largest and most diverse dataset of street scenes with high-quality and coarse annotations at the time of publication; (ii) developing a sound evaluation methodology for pixel-level and instance-level semantic labelling; (iii) providing an in-depth analysis of the characteristics of our dataset; (iv) evaluating several state-of-the-art approaches on our benchmark.

One key observation from our analysis is that the relative order of performance for the state-of-the-art on our dataset is notably different than on more generic datasets such as *PASCAL VOC*. Our conclusion is that serious progress in urban scene understanding may not be achievable through such generic datasets, as the latter (exemplified by *Cityscapes*) pose unique challenges, such as highly crowded scenes, difficult imaging conditions due to motion blur and contrast variation, as well as a large variance in object scale.

At publication time, the best-performing baseline for pixel-level semantic segmentation obtains an IoU score of 67.1 %. 144 entries later on the public benchmark table, the current record stands at 83.6 %. The instance-level task has proven to be particularly challenging with an mAP^r score of 38.0 %. For comparison: The best current method on the *MSCOCO* Lin *et al.* (2014) instance-level segmentation benchmark attains an mAP^r score of 49.0 %.

Since publication of the dataset, several large-scale street scene understanding datasets have been released, e.g. *Mapillary Vistas* (Neuhold *et al.*, 2017), *Berkeley DeepDrive* (Yu *et al.*, 2020), *Apolloscape* (Huang *et al.*, 2018). While these exceed Cityscapes in terms of size and image diversity, the former remains the de facto standard benchmark for pixel-level and instance-level street scene understanding, that is far from solved especially when it comes to the latter task. It has also found further uses beyond the intended ones, e.g. for the generative modelling and synthesis of images (Wang *et al.*, 2018b; Zhu *et al.*, 2017b; Liu *et al.*, 2017), and has been extended with new annotations (Zhang *et al.*, 2017b) and imagery (Sakaridis *et al.*, 2018), with further extensions targeting more fine-grained annotations in the pipeline.

8 Weakly-Supervised Boundary Detection

As seen in previous chapters, deep learning-based recognition methods benefit from — if not outright require — large amounts of annotated data. Obtaining this data is more feasible for some tasks (e.g. detection, semantic labelling) than others. Boundary detection is an instance of the latter set of problems, for which there is a need to relax the requirement to carefully annotate images to make both the training more affordable and to extend the amount of training data.

In this chapter we propose a technique to generate weakly supervised annotations and show that bounding box annotations alone suffice to reach high-quality object boundaries without using any object-specific boundary annotations. With the proposed weak supervision techniques we achieve the top performance on the object boundary detection task, outperforming by a large margin the current strongly supervised state-of-the-art methods.

This work was published at CVPR (Khoreva *et al.*, 2016). Anna Khoreva was the lead author and Mohamed Omran conducted all the experiments involving neural networks.

8.1 Introduction

Boundary detection is a classic computer vision problem. It is an enabling ingredient for many vision tasks such as image/video segmentation (Arbelaez *et al.*, 2011; Galasso *et al.*, 2013), object detection (Hosang *et al.*, 2016; Zhu *et al.*, 2015), and semantic labelling (Banica and Sminchisescu, 2015). What constitutes a boundary in the image is task-dependent. In the context of the aforementioned tasks, boundaries are taken to mean the edges that separate objects from the background or from other objects. As these tasks typically target a pre-defined set of classes, we are accordingly interested in detecting the boundaries of objects from these classes. In this chapter, we address the class-specific (or semantic) boundary detection problem.

State-of-the-art boundary detection relies on learning-based methods which in turn require extensive training data. However, instance-wise boundary annotations are very expensive to obtain. Compared to two clicks for a bounding box, annotating the boundary of an object often requires drawing a polygon with 20~100 clicks, i.e. an increase in effort of 1-2 orders of magnitude.

To better train deep models and extend their coverage to more object classes, there is a need to relax the requirement of high-quality image annotations. Our starting point

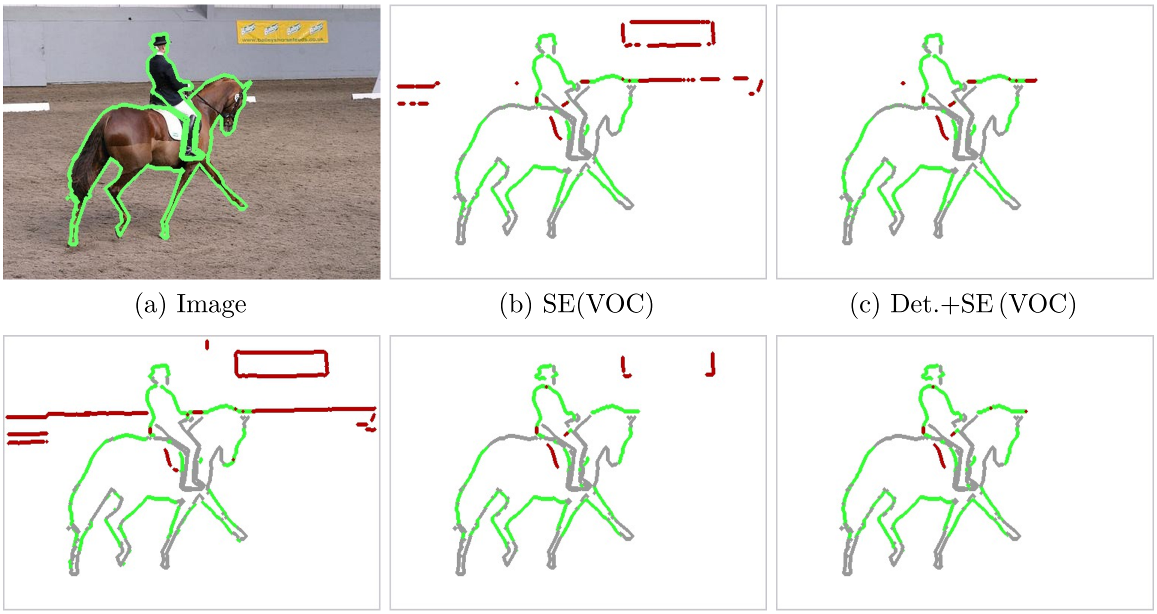

(a) Image
(b) SE(VOC)
(c) Det.+SE (VOC)
(d) SE(BSDS)
(e) SE (weak)
(f) Det.+SE (weak)

Figure 8.1: Object-specific boundaries (a) differ from generic boundaries such as the ones detected in (d). The proposed weakly supervised approach drives boundary detection towards the objects of interest. Example results in (e) and (f). Red/green indicate false/true positive pixels, grey are undetected boundary pixels. All methods shown at 50% recall.

in this chapter is thus the following question: Can we obtain reliable object-specific boundaries without having access to object boundary annotations at training time?

In this chapter we focus on learning object boundaries in a weakly supervised fashion and show that high quality object boundary detections can be obtained *without* using any class-specific boundary annotations. We propose several ways of generating object boundary annotations with different levels of supervision: either with a boundary detector trained on generic boundary annotations (from the *BSDS500* dataset), as well as just using a bounding-box-based object detector. In the latter case, we generate weak object boundary annotations by combining unsupervised image segmentation (Felzenszwalb and Huttenlocher, 2004), region-based object proposal methods (Uijlings *et al.*, 2013; Pont-Tuset *et al.*, 2017) and object detectors (Girshick, 2015; Ren *et al.*, 2015). We show that with bounding box annotations alone, we can obtain high quality object boundary estimates.

We present results using a decision forest (Dollár and Zitnick, 2015) and a CNN-based edge detector (Xie and Tu, 2017). We report top performance on the *PASCAL VOC2012* object boundary detection benchmark (*SBD*) (Hariharan *et al.*, 2011; Everingham *et al.*, 2015) with our weakly supervised approaches, already surpassing previously reported strongly supervised results.

Our main contributions are summarised below:

- We introduce the problem of weakly supervised object-specific boundary detection.

- We show that good boundary estimates can be obtained on *BSDS500*, *PASCAL VOC2012*, and *SBD* using only weak supervision, namely by leveraging bounding box detection annotations without the need for instance-wise object boundary annotations.

- We report the best known results on *PASCAL VOC2012* and *SBD*. Our weakly supervised results alone improve over the previous strongly supervised state-of-the-art.

The rest of this chapter is organised as follows. In Sec. 8.2, we introduce some related work on boundary detection and weakly-supervised learning. Sec. 8.3 describes different types of boundary detection and the relevant datasets we consider here. In Sec. 8.4, we proposal several approaches for generating boundary annotations with varying levels of supervision. We then experiment with these schemes and report the results for the different boundary detection tasks in the remaining sections.

8.2 Related work

Generic boundaries Early methods in boundary detection rely on a fixed prior model of what constitutes a boundary, e.g. the *Canny* detector (Canny, 1986). Modern methods resort to data-driven techniques that learn to predict if a pixel belongs to a boundary. From well-crafted features and simple classifiers, e.g. *gPb* (Arbelaez *et al.*, 2011), to powerful decision trees over fixed features, e.g. *SE* (Dollár and Zitnick, 2015) and *OEF* (Hallman and Fowlkes, 2015), and recently to end-to-end learning via CNNs, e.g *DeepEdge* (Bertasius *et al.*, 2015), *N4* (Ganin and Lempitsky, 2014), and *HED* (Xie and Tu, 2017). CNNs are usually pre-trained on large classification datasets, so as to be initialised with reasonable features. The more sophisticated the model, the more data is required to learn it.

As an alternative to direct boundary prediction, segmentation techniques can also be used to improve boundary estimates or to generate closed contours, e.g. *F&H* (Felzenszwalb and Huttenlocher, 2004), *gPb-owt-ucm* (Arbelaez *et al.*, 2011), and *MCG* (Pont-Tuset *et al.*, 2017).

While the overwhelming majority of recent methods rely on supervised learning, a few recent works have addressed unsupervised detection of generic boundaries (Isola *et al.*, 2014; Li *et al.*, 2016). *PMI* (Isola *et al.*, 2014) detects boundaries by modelling them as statistical anomalies amongst all local image patches, reaching competitive performance without the need for learning. Recently, Li *et al.* (2016) propose to train edge detectors using motion boundaries obtained from a large corpus of video data instead of resorting to manually annotated images. Both approaches attain similar detection performance.

Object-specific boundaries In many applications, there is interest in boundaries of specific object classes. The class-specific object boundary detectors need then to be trained or tuned to the classes of interest. This problem is more recent and still relatively unexplored. Hariharan *et al.* (2011) introduced the *SBD* dataset to measure this task over the 20 *PASCAL VOC* categories (Everingham *et al.*, 2015). Additionally, they present a method that re-weights generic boundaries using the activation regions of a detector. Uijlings and Ferrari (2015) propose to train class-specific boundary detectors, and weigh them at test time according to an image classifier.

Weakly supervised learning In this work we are interested in object-specific boundaries *without using class-specific boundary annotations.* We only use bounding box annotations, and in some experiments, generic boundaries from the *BSDS500* dataset (Arbelaez *et al.*, 2011). Multiple works have addressed weakly supervised learning for object localisation (Oquab *et al.*, 2015; Cao *et al.*, 2015), object detection (Prest *et al.*, 2012; Wang *et al.*, 2014a), or semantic labelling (Vezhnevets *et al.*, 2011; Xu *et al.*, 2015; Pinheiro and Collobert, 2015). To the best of our knowledge, there is no previous work attempting to learn object boundaries in a weakly supervised fashion.

8.3 Setting: Tasks, Datasets & Baselines

8.3.1 Tasks

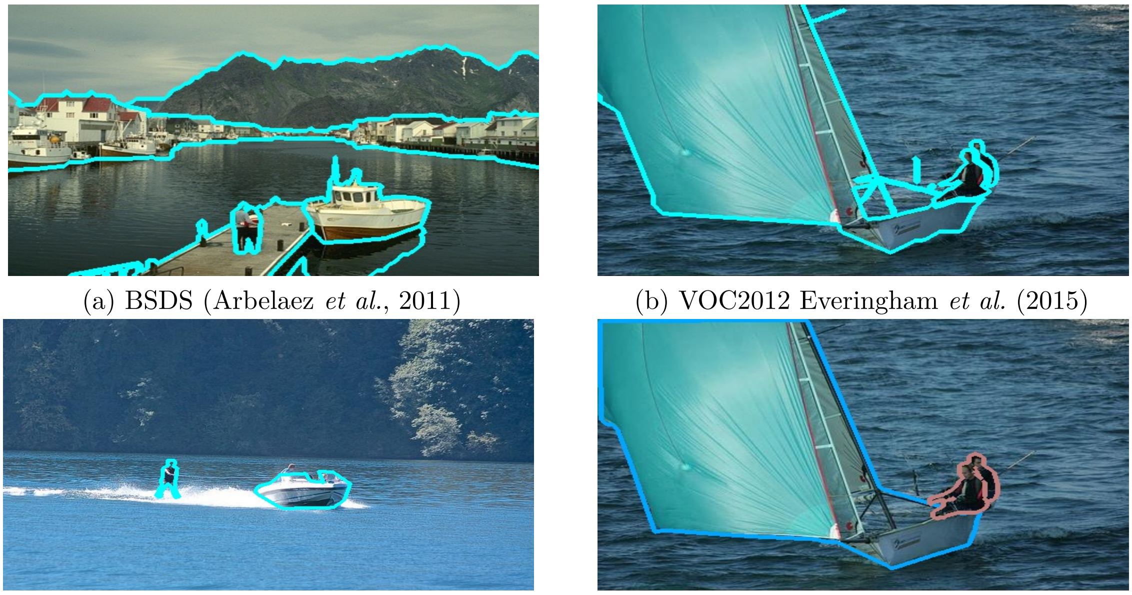
In this work we distinguish between three types of boundaries: (i) generic boundaries (delineating both "things" and "stuff", as well as salient surface and texture boundaries), (ii) instance-wise boundaries (external object instance boundaries), and (iii) class specific boundaries (object instance boundaries of a certain semantic class).

8.3.2 Datasets

For detecting these three types of boundaries we consider different datasets: the Berkeley Segmentation Dataset and Benchmark (*BSDS500* or hereafter: *BSDS*) (Martin *et al.*, 2001; Arbelaez *et al.*, 2011), *PASCAL VOC2012* (*VOC*) (Everingham *et al.*, 2015), *MSCOCO* (*COCO*) (Lin *et al.*, 2014), and the *Semantic Boundary Dataset* (*SBD*) (Hariharan *et al.*, 2011), where each represents boundary annotations of a given boundary type (see Fig. 8.2).

BSDS We first present our results on the *BSDS*, the most established benchmark for generic boundary detection. The dataset contains 200 training, 100 validation and 200 test images. Each image has multiple ground truth annotations. To evaluate the quality of estimated boundaries three measures are used: fixed contour threshold (ODS), per-image best threshold (OIS), and average precision (AP). Following the standard approach, prior to evaluation we apply a non-maximal suppression technique to boundary probability maps to obtain thinned edges (Dollár and Zitnick, 2015; Canny, 1986).

VOC For evaluating instance-wise boundaries we propose to use the *VOC* segmentation dataset. The dataset contains 1 464 training and 1 449 validation images, annotated



(a) BSDS (Arbelaez *et al.*, 2011)

(b) VOC2012 Everingham *et al.* (2015)

(c) COCO Lin *et al.* (2014)

(d) SBD Hariharan *et al.* (2011)

Figure 8.2: Datasets considered.

with contours for 20 object classes for all instances. The dataset was originally designed for semantic segmentation. Therefore only object interior pixels are marked and the boundary location is recovered from the segmentation mask. Here we consider only object boundaries without distinguishing between different classes, treating all 20 classes as one. For measuring the quality of predicted boundaries the *BSDS* evaluation software is used. Following Uijlings and Ferrari (2015) the *maxDist* parameter (maximum tolerance for edge matches) is set to 0.01.

Since we generate boundary annotations in a weakly supervised fashion, we are able to generate boundaries over arbitrary image sets. Besides the *VOC* segmentation dataset, we can also use images from the *VOC* detection set. The combination of the two is referred to here as VOC_{+}.

COCO As an additional benchmark for instance-wise boundary detection, we use *COCO*. The dataset provides semantic segmentation masks for 80 object classes. For our experiments we consider only images that contain the 20 *VOC* classes and objects larger than 200 pixels. The subset of *COCO* images that contain *VOC* classes consists of 65 813 training and 30 163 validation images. For computational reasons we limit evaluation to 5 000 randomly chosen images of the validation set. The default settings of the *BSDS* evaluation software is used ($maxDist = 0.01$). Only object boundaries are evaluated without requiring these to be assigned class labels.

SBD We use *SBD* for evaluating class-specific object boundary predictions. The dataset consists of 11 318 images from the *trainval* set of the *PASCAL VOC2011* challenge, divided into 8 498 training and 2 820 test images. This dataset has object instance boundaries with accurate figure/ground masks that are also labelled with one of 20 *VOC* classes. The boundary detection accuracy for each class is evaluated using the official evaluation software (Hariharan *et al.*, 2011). During the evaluation process all internal object-specific boundaries are set to zero and the *maxDist* parameter is set to 0.02. We report the mean ODS F-measure (F), and average precision (AP) across 20 classes.

Note that there are overlaps between the training and test sets of *VOC* and *SBD*. For cross-dataset experiments we make sure not to re-use any images included in the test set considered.

8.3.3 Baselines

For our experiments we consider two different types of boundary detectors as baselines: *SE* (Dollár and Zitnick, 2015) and *HED* (Xie and Tu, 2017).

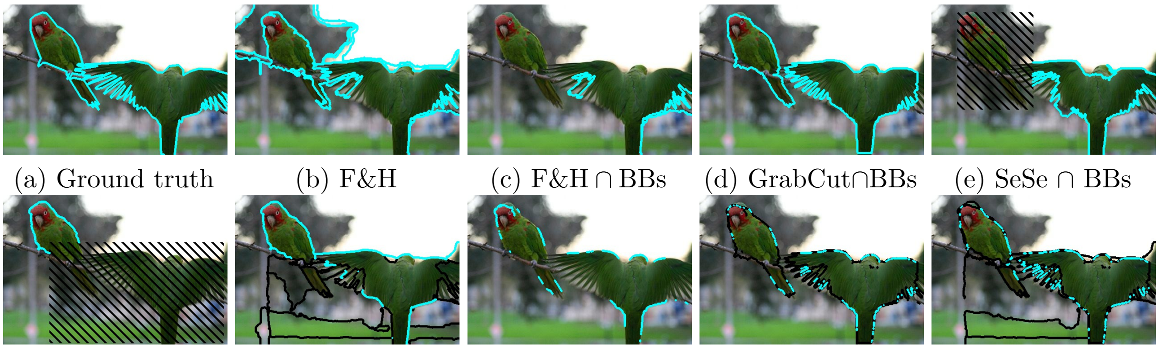
SE is at the core of multiple related methods, e.g. *SCG* (Ren and Bo, 2012), *MCG* (Pont-Tuset *et al.*, 2017), and *OEF* (Hallman and Fowlkes, 2015). *SE* builds a "structured decision forest" which is a modified decision forest, where the leaf outputs are local boundary patches (16 × 16 pixels) as opposed to single pixel-wise predictions. The patch-wise outputs are averaged at test time, and the split nodes are built taking into account the local segmentation of the ground truth input patches. The split decision function uses binary comparisons, selecting among hand-crafted edge and self-similarity features. This method requires closed contours (i.e. segmentations) as training inputs. This detector is reasonably fast to train/test and yields good detection quality.

HED is currently the top performing CNN for *BSDS* boundaries. It builds upon a *VGG16* network (Simonyan and Zisserman, 2015) pre-trained on *ImageNet* (Russakovsky *et al.*, 2015a), and exploits features from all layers to build its output boundary probability map. By also exploiting the lower layers (which have higher resolution) the output is more detailed, and the fine-tuning is more effective (since all layers are guided directly towards the boundary detection task). To reach top performance, *HED* is trained using a subset of the annotated *BSDS* pixels, on which annotators agree. These are so called "consensus" annotations (Hou *et al.*, 2013), and correspond to ∼15% of all true positives.

8.4 Methods: Generating Weak Supervision

Our goal is to generate (noisy) training data for boundary detectors to reduce the annotation burden. In this section we describe several approaches to generating such weak supervision. Different combinations of methods will be applicable depending on whether we target generic or class-specific (semantic) boundaries. Some of the approaches we consider are illustrated in Fig. 8.3.

BBs For class-specific boundary detection, as the classes of interest are specified in advance, we can use an object detector to filter generic boundary detections to generate



(a) Ground truth (b) F&H (c) F&H ∩ BBs (d) GrabCut∩BBs (e) SeSe ∩ BBs

(f) MCG ∩ BBs (g) cons.MCG∩BBs (h) SE(SeSe ∩ BBs) (i) cons.S&G∩BBs (j) cons.ALL ∩ BBs

Figure 8.3: Different generated boundary annotations. Cyan/black indicates positive/ignored boundaries.

the required training data. This potentially alleviates the need for class-specific boundary annotations. To this end, we use the *Fast R-CNN* detector (Girshick, 2015), which for training only requires weak annotations in the form of bounding boxes. We apply this detector to the training set (and possibly a larger set of images), and retain boxes with confidence scores above 0.8. We also experimented with using the ground truth annotations directly, but saw no noticeable difference. We thus report numbers only using the "detections over the training set".

F&H As a source of unsupervised boundaries we consider *F&H*, the classical graph-based image segmentation technique proposed by Felzenszwalb and Huttenlocher (2004) (Fig. 8.3b). We use this directly as a form of weak supervision for the generic boundary detection task in Sec. 8.5. As described above, we additionally use the bounding boxes produced by an object detector to focus the resulting data on classes of interest. This combination is referred to as *F&H ∩ BBs*. Only the boundaries of segments that are contained inside a bounding box are retained. Poorly-aligned detections can thus result in missed boundaries as can be seen in Fig. 8.3c.

GrabCut Boundaries from *F&H* will trigger on any kind of edge, including the internal edges of objects that results from texture or surface changes. One way to exclude internal edges is to extract the external boundaries via figure-ground segmentation of the sub-image enclosed within a bounding box. We use *GrabCut* (Rother *et al.*, 2004) for this purpose. We also experimented with *DenseCut* (Cheng *et al.*, 2015), but did not obtain any gains and thus we only report results for *GrabCut∩BBs* (Fig. 8.3d). *GrabCut* might result in inaccurate segments, so we reject a segment if it has an intersection-over-union score (IoU) ≥ 0.7 with the corresponding bounding box. In this case, the area of the bounding box is marked as an ignore region (see Fig. 8.3e for an example).

Object proposals Another way to bias generation of boundary annotations towards object contours is to consider object proposal methods. *Selective Search* (*SeSe*) (Uijlings *et al.*, 2013) is based on *F&H* segmentations — thus is fully unsupervised, while *MCG* (Pont-Tuset *et al.*, 2017) employs boundaries estimated via *SE*(*BSDS*) — thus uses generic boundary annotations.

As with *GrabCut*$\cap$*BBs*, we generate *SeSe*$\cap$*BBs* (Fig. 8.3e) and *MCG*$\cap$*BBs* (Fig. 8.3f) by matching proposals to bounding boxes based on IoU. We use a stricter threshold of IoU ≥ 0.9 as object proposals tend to be better localised. When more than one proposal is matched to a detection bounding box we use the union of the proposal boundaries as positive annotations. This maximises boundary recall, and somewhat imitates the *BSDS* annotation protocol which involves multiple human annotators. We also experimented with using only the highest overlapping proposal, but the union provides marginally better results; thus we report only results with the union. Once we select a proposal, we do not exclude boundaries that lie outside the bounding box as the latter might not be well-aligned with the object.

Consensus boundaries Due to the ill-defined nature of boundaries, different human annotators will produce different boundary annotations. Some methods thus resort to selecting as positive training examples boundaries on which there is agreement among annotators. In fact as we will demonstrate later in Tab. 8.1, *HED* requires such consensus boundaries to reach good performance.

Thus rather than taking the union between proposal boundaries, we consider using the consensus between object proposal boundaries. The boundary is considered to be present if the agreement is higher than 70%, otherwise the boundary is ignored. We denote such generated annotations as "cons.", e.g. *cons.MCG*$\cap$ *BBs* (Fig. 8.3g).

Another way to generate sparse, consensus-like boundaries, is to threshold the boundary probability map out of an *SE*$(\cdot)$ model. *SE*(*SeSe*$\cap$ *BBs*) uses the top 15% quantile per image as weakly supervised annotations (Fig. 8.3h).

Finally, besides the consensus between proposals, we can also rely on the consensus between methods. *cons.S&G*$\cap$*BBs* (Fig. 8.3i) is the intersection between *SE*(*SeSe*$\cap$ *BBs*), *SeSe* and *GrabCut* (fully unsupervised), while *cons.ALL*$\cap$ *BBs* (Fig. 8.3j) is the intersection between *MCG*, *SeSe* and *GrabCut* (uses *BSDS* data).

8.5 Results: Generic Boundary Detection

We start by exploring weakly supervised training for generic boundary detection, i.e. the task specified by *BSDS*. In the case of this task, weak supervision means deriving targets from learning-free boundary detectors such as *Canny* (Canny, 1986) and *F&H* (Felzenszwalb and Huttenlocher, 2004), which provide relatively low quality detections. We notice that correct boundaries tend to have consistent appearance, while boundaries with inconsistent appearance more often result in erroneous detections. Robust training methods should however be able to pick up the signal in such noisy detections. In Fig. 8.4 and Tab. 8.1 we report our results when training a structured decision forest (*SE*) and a CNN (*HED*) with noisy boundary annotations. By $(\cdot)$ we denote the data used for training.

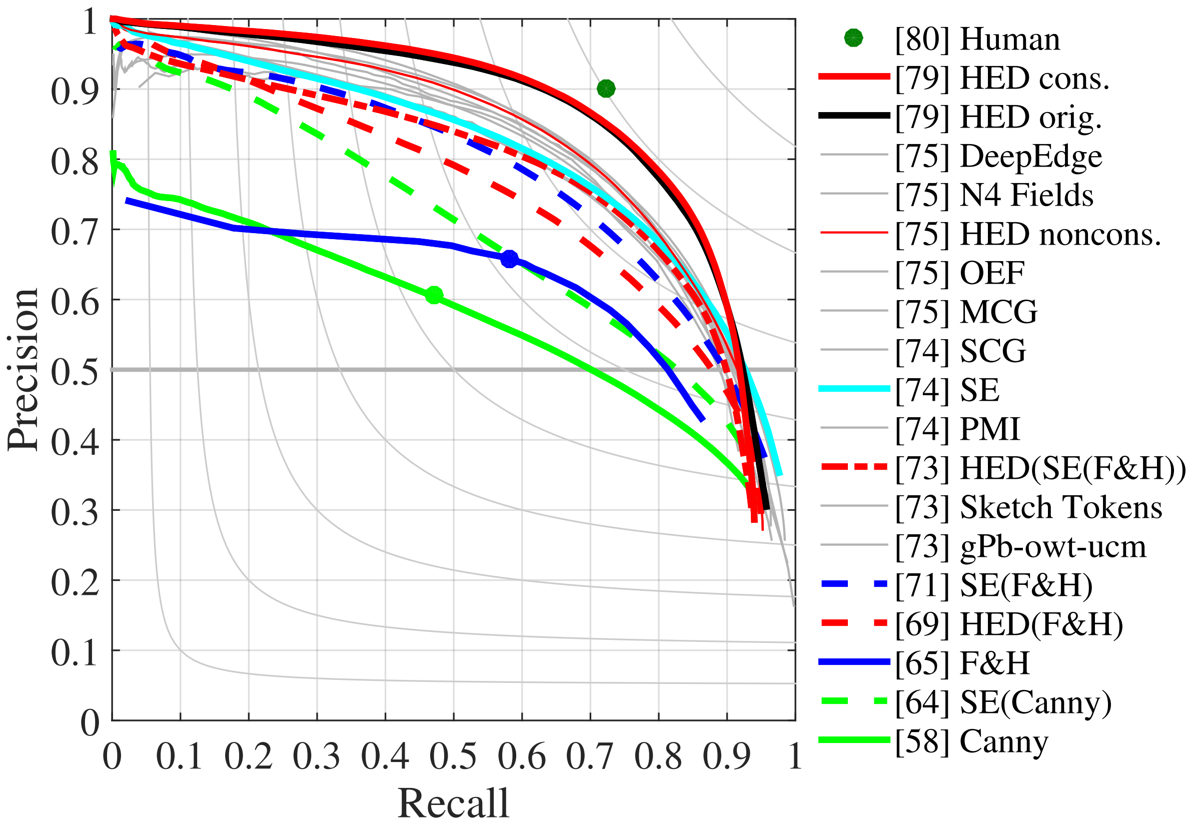


Figure 8.4: *BSDS* results. *Canny* and *F&H* points indicate the boundaries used as noisy annotations. When trained over noisy annotations, both *SE* and *HED* provide a large quality improvement.

SE When training *SE* either using *Canny* — *SE*(*Canny*)) — or using *F&H* — *SE*(*F&H*)) — we observe a notable jump in boundary detection quality. *SE*(*F&H*) closes up to 80% of the gap between *SE* trained with the BSDS ground truth — *SE*(*BSDS*) (strong supervision) — and *F&H* (ΔAP% column in Tab. 8.1). Using only noisy weak supervision *SE*(*F&H*) is only 3 points behind the strongly supervised case (76 vs. 79).

We believe that the strong noise robustness of *SE* can be attributed to the way it builds its leaves. The final output of each leaf is the medoid of all segments reaching it. If the noisy boundaries are randomly spread in the image appearance space, the medoid selection will be robust.

HED *HED* reaches top quality when trained over consensus annotations. When using all annotations (“non-consensus”), its performance is comparable to other CNN-based alternatives. When *HED* is trained with annotations derived from *F&H*, the relative improvement is smaller than the corresponding improvement for *SE*. When combined with *SE* (denoted *HED(SE*(*F&H*))) it reaches 69 ΔAP%. This two-stage approach provides better boundaries than *SE*(*F&H*) alone, and reaches a quality comparable to the classic *gPb* method Arbelaez *et al.* (2011) (75 vs. 73).

Family	Method	ODS	OIS	AP	ΔAP%
Unsupervised	Canny	58	62	55	-
	F&H	64	67	64	-
	PMI	74	77	78	-
Trained on ground truth	gPb-owt-ucm	73	76	73	-
	SE(BSDS)	74	76	79	-
	HED(BSDS) noncons.	75	77	80	-
	HED(BSDS) cons.	79	81	84	-
Trained on unsupervised boundary estimates	SE (Canny)	64	67	64	38
	SE (F&H)	71	74	**76**	80
	SE (SE (F&H))	72	74	76	80
	SE(PMI)	72	75	77	-
	HED (F&H)	69	72	73	56
	HED (SE (F&H))	73	76	**75**	69

Table 8.1: Detailed *BSDS* results, see Fig. 8.4 and Sec. 8.5. Underlined results correspond to baselines that rely on ground truth boundaries, and our best weakly supervised results are in boldface. $(\cdot)$ denotes the data used for training. ΔAP% indicates the ratio between the same model trained on ground truth, and the noisy input boundaries. The closer to 100%, the lower the drop due to using noisy inputs instead of ground truth.

We should note that on *BSDS*, the unsupervised *PMI* method provides better boundaries than our weakly supervised variants. However *PMI* cannot be adapted to provide object-specific boundaries. For this we need to rely on methods than can be trained with class-specific annotations, such as *SE* and *HED*.

Conclusion *SE* is surprisingly robust to annotation noise during training. *HED* is also robust but to a lesser degree. By using noisy boundaries generated from unsupervised methods, we can reach a performance competitive with recent strongly supervised methods.

8.6 Results: Class-specific Boundary Detection

In this section we analyse the variants of weakly supervised methods for object boundary detection proposed in Sec. 8.4. Here, we're interested in detecting the boundaries of object instances belonging to a specific set of classes, but not in predicting the class labels themselves. This means that all 20 *VOC* classes are treated as a single class. The evaluation protocol is described in Sec. 8.3.1. Since we're interested in the boundaries for specific objects, we take weak supervision to mean generic boundary annotations as well as bounding box-level object annotations. We use the bounding box annotations in two ways: (i) to filter out noisy boundaries for the training data, but also (ii) to post-process boundary predictions by upweighting ones that coincide with object detections and downweighting the others. We will describe the exact process shortly. First, we discuss results using *SE* then results with *HED*.

8.6.1 Structured Forests (VOC)

We'll start by establishing different baseline results using strong supervision, and then compare these against models trained with different variants of weak supervision.

8.6.1.1 Strong Supervision

SE Fig. 8.5a and Tab. 8.2 show results of *SE* trained over the ground truth of different datasets (dashed lines). Our results of *SE*(*VOC*) are on par with the ones reported in Uijlings and Ferrari (2015). The gap between SE (VOC) and SE (BSDS) reflects the difference between generic boundaries and boundaries specific to the 20 *VOC* object categories (see also Fig. 9.1).

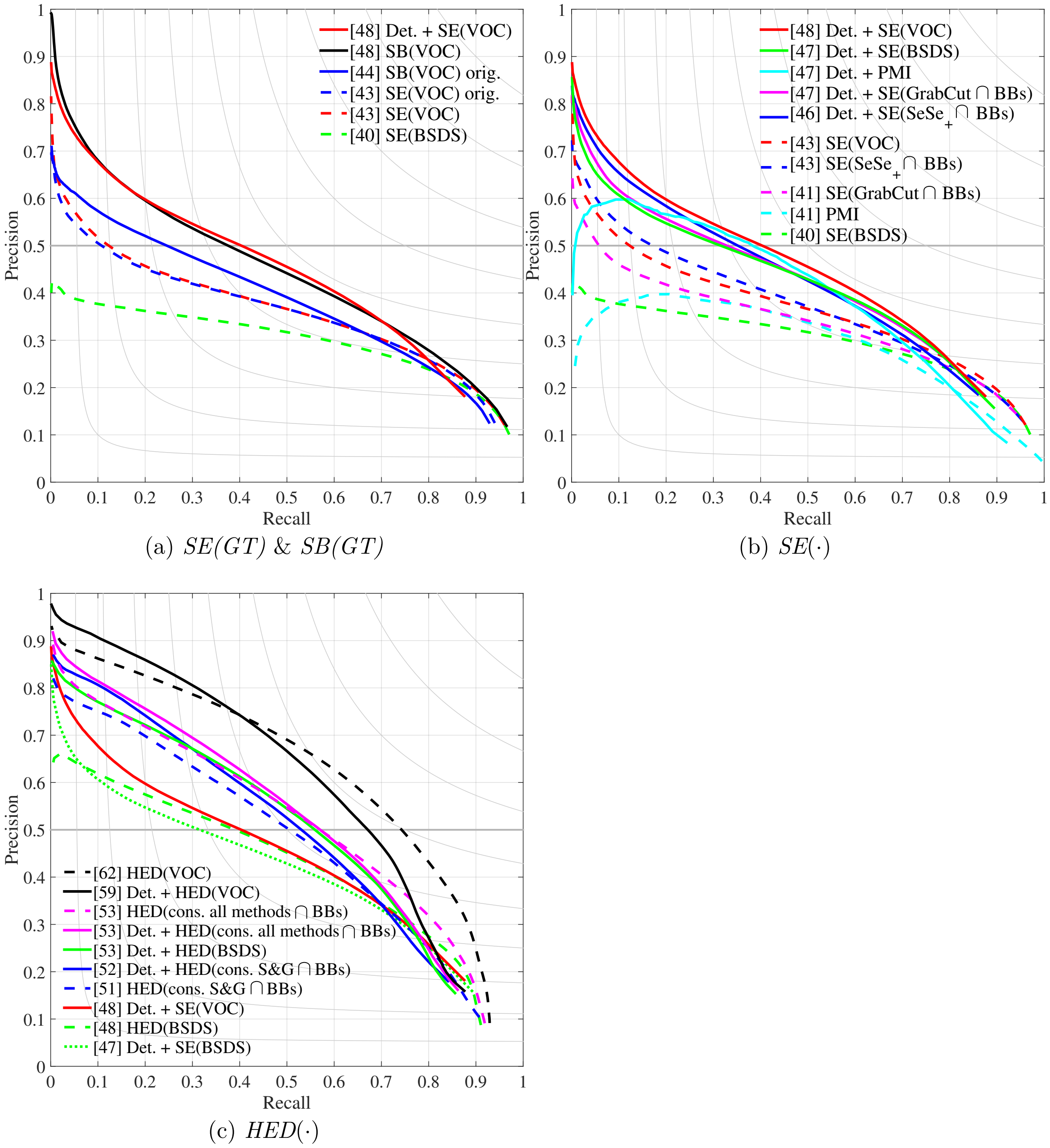
SB To improve object-specific boundary detection, the Situational Object Boundary detector (*SB*) (Uijlings and Ferrari, 2015), trains 20 class-specific *SE* models. These models are combined at test time using a CNN-based image classifier. The original *SB* results as well as our reproduction of these results *SB*(*VOC*) are shown in Fig. 8.5a. Our version obtains better results (+4AP) due to training the *SE* models with more samples per image, and using a stronger CNN (Simonyan and Zisserman, 2015).

Detector + SE Rather than training and testing 20 *SE* models plus an image classifier, we propose to leverage the same training data using a single *SE* model together with a detector (Girshick, 2015). By computing a per-pixel maximum among all detection bounding boxes and their score, we construct an "objectness map" that we multiply with the boundary probability map from *SE*. False positive boundaries are thus downweighted, and boundaries in high confidence regions for the detector are boosted. The detector is trained with the same per-object boundary annotations used to train the *SE* model, no additional data is required.

Our *Det.+SE*(*VOC*) obtains the same detection quality as *SB*(*VOC*) while using only a single *SE* model. These are the best reported results on this task (top of Tab. 8.2), when using strong supervision. One could in principle also combine object detection with *SB* for even stronger results, but we leave this for future work.

8.6.1.2 Weak Supervision

Given the reference performance of *Det.+SE*(*VOC*), can we reach similar boundary detection quality without using the boundary annotations from *VOC*?



(a) *SE(GT)* & *SB(GT)*

(b) *SE*(·)

(c) *HED*(·)

Figure 8.5: *VOC* results for (a) strongly- and (b) weakly-supervised *SE*-based models (Sec. 8.6.1), as well as for (c) weakly-supervised *HED* models (Sec. 8.6.2). (·) indicates the data used for training. Curves with continuous lines correspond to models that rely on an additional CNN-based classifier or detector at test time, and dashed lines correspond to models that don't. The curves are summarised in the legend with AP. The modifier "orig." indicates original results from Dollár and Zitnick (2015) and Uijlings and Ferrari (2015) respectively, which we also reproduce ourselves here. Det. indicates results that involve post-processing with object detections (see Sec. 8.6.1).

Family	Method	Data	Without BBs			With BBs		
			F	AP	ΔAP	F	AP	ΔAP
GT	SE	VOC	43	35	-	48	41	-
Other GT	SE	COCO	44	37	2	49	42	1
	SE	BSDS	40	29	-6	47	39	-2
	MCG		41	28	-7	48	39	-2
Weakly supervised	SE	F&H ∩ BBs	40	29	-6	46	36	-5
		GrabCut ∩ BBs	41	32	-3	47	39	-2
		SeSe ∩ BBs	42	35	0	46	39	-2
		SeSe$_+$ ∩ BBs	**43**	**36**	**+1**	46	39	-2
		MCG ∩ BBs	43	34	-1	47	39	-2
		MCG$_+$ ∩ BBs	43	35	0	**48**	**40**	**-1**
Unsupervised	F&H	-	34	15	-20	41	25	-16
	PMI		41	29	-6	47	38	-3

Table 8.2: *VOC* results for *SE* models, see Figs. 8.5a and 8.5b for the full curves. Underlined results correspond to baselines that rely on ground truth boundaries, and our best weakly supervised results are in boldface.

SE $(\cdot)$ An *SE* model trained using the *BSDS* annotations attains relatively low performance (see *SE*(*BSDS*) in Fig. 8.5b), as does *PMI*. The same *BSDS* data can be used to generate *MCG* object proposals for the *VOC* training data, and a detector trained on *VOC* bounding boxes can generate bounding boxes for the same data. We combine these to generate boundary annotations (*MCG* ∩ *BBs*) as described in Sec. 8.4. These lead to improved results over the *BSDS*-trained baseline. By extending the training set to the additional *VOC*$_+$ images (*SE*(*MCG*$_+$ ∩ *BBs*) in Tab. 8.2) we match the performance of a strongly-supervised *SE* model (*SE*(*VOC*)). We also consider variants that do not require the *BSDS* ground truth, such as *SeSe* and *GrabCut*. *SeSe*-derived boundaries lead to essentially the same results as data obtained with *MCG*.

Det.+SE $(\cdot)$ Post-processing the results at test time with an object detector as previously described minimises the differences between all weakly supervised methods. *Det.+PMI* shows strong results, but (since *PMI* is learning-free) fails to reach high precision. The high quality of *Det.+BSDS* indicates that *BSDS* annotations, despite being in principle “generic boundaries” reflect object boundaries well, at least in the proximity of an object. This is further confirmed in Sec. 8.6.2. Compared to *Det.+BSDS* our weakly supervised annotation further close the gap to *Det.+SE*(*VOC*) (especially in the high precision regime), even when not using any *BSDS* data.

Conclusion By using bounding box annotations via an object detector, our weakly supervised boundary annotations enable the *Det.+SE* model to match the strongly supervised model, improving over the best reported results on the task. We also observe that *BSDS* data allows us to train models that detect object boundaries well.

Family	Method	Data	Without BBs			With BBs		
			F	AP	ΔAP	F	AP	ΔAP
GT	SE	VOC	43	35	-	48	41	-
	HED		62	61	26	59	58	17
Other GT	HED	BSDS	48	41	6	53	48	7
		COCO	59	60	25	56	55	14
Weakly supervised	SE	MCG ∩ BBs	43	34	-1	47	39	-2
	HED	SE(SeSe ∩ BBs)	45	37	3	49	40	-1
		MCG ∩ BBs	50	44	9	48	42	1
		cons. S&G ∩ BBs	**51**	**46**	**+11**	**52**	**47**	**+8**
		cons. MCG ∩ BBs	53	50	15	52	49	8
		cons.ALL∩BBs	**53**	**50**	**+15**	**53**	**50**	**+9**

Table 8.3: VOC results for *HED* models, see Fig. 8.5c. Underlined results correspond to baselines that rely on ground truth boundaries, and our best weakly supervised results are in boldface.

8.6.2 CNNs (VOC)

This section analyses the performance of *HED* (Xie and Tu, 2017) trained with the weakly supervised variants proposed in Sec. 8.4. We use our re-implementation of *HED* which performs on par with the original (see Fig. 8.4). We use the same evaluation setup as in the previous section. Fig. 8.5c and Tab. 8.3 show the results.

HED (·) *HED(VOC)* outperforms *SE(VOC)* by a large margin. By comparing their predictions qualitatively, we observe that *HED* manages to suppress the internal object boundaries well, while *SE* fails to do so probably due to its decisions being based on more local support, whereas *HED* incorporates more context.

HED(BSDS) achieves high performance on the object boundary detection task, despite being trained with generic boundaries. Specifically, *HED(BSDS)* is trained on “consensus” annotations which are closer to object-like boundaries: The fraction of annotators agreeing on the presence of external object boundaries is much higher than for non-object or internal object boundaries.

For training *HED*, in contrast to the *SE* model, we do not need closed contours and can use the consensus between different weak annotation variants. This results in better performance. Using the consensus between boundaries of *MCG* proposals *HED*(*cons.MCG* ∩ *BBs*) improves AP by 6% compared to using the union of object proposals *HED*(*MCG* ∩ *BBs*) (see Tab. 8.3).

The *HED* models trained with weak annotations outperform the fully supervised *SE(VOC*) and do not reach the performance of *HED(VOC)*. As has been shown in Sec. 8.5 the *HED* detector is less robust to noise than *SE*.

Method	Family	Data	Without BBs			With BBs		
			F	AP	ΔAP	F	AP	ΔAP
SE	GT	COCO	$\underline{40}$	$\underline{32}$	-	$\underline{45}$	$\underline{37}$	-
	Other GT	BSDS	34	23	-9	43	33	-4
	Weakly supervised	$\text{SeSe}_+ \cap$ BBs	**40**	**31**	**-1**	**44**	**35**	**-2**
		$\text{MCG}_+ \cap$ BBs	39	30	-2	44	35	-2
HED	GT	COCO	60	59	27	56	55	18
	Other GT	BSDS	44	34	2	49	42	5
	Weakly supervised	cons. S&G∩BBs	47	39	7	48	42	5
		cons.ALL∩BBs	**49**	**43**	**+11**	**50**	**44**	**+7**

Table 8.4: COCO results. Underlined results correspond to baselines that rely on ground truth boundaries.

Det.+HED (·) Combining an object detector with HED(VOC) (see Det.+HED (VOC) in Fig. 8.5c) is not beneficial to the performance as the *HED* detector already has notion of objects and their location due to pixel-to-pixel end-to-end learning of the network.

For *HED* models trained with the weakly supervised variants, employing an object detector at test time brings only a slight improvement of the performance in the high precision area. The reason for this is that we already use information from the bounding box detector to generate the annotation and the CNN-based method is able to learn it during training.

Det.+HED (MCG ∩ BBs) outperforms *Det.+HED (BSDS)* (see Tab. 8.3). Note that the *HED* trained with the proposed annotations, generated without using boundary ground truth, performs on par with the *HED* model trained on generic boundaries (*Det.+HED (cons. S&G∩BBs)* and *Det.+HED (BSDS)* in Fig. 8.5c).

The qualitative results are presented in Fig. 8.6 and provide support for the quantitative evaluation.

Conclusion Similar to other computer vision tasks CNN-based methods show superior performance compared to more traditional approaches. Due to the pixel-to-pixel training and global view of the image CNNs seem to have a notion of objects and their locations which allows us to omit the use of the detector at test time. With our weakly supervised boundary annotations we can gain fair performance without using any instance-wise object boundary or generic boundary annotations. We leave out object detection at test time, and only feed object bounding box information during training.

8.6.3 Further Results (COCO)

Additionally we show the generalisation of the proposed weakly supervised variants for object boundary detection on *COCO*. We use the same evaluation protocol as for *VOC*. For weakly supervised cases the results are shown with the models trained on *VOC*

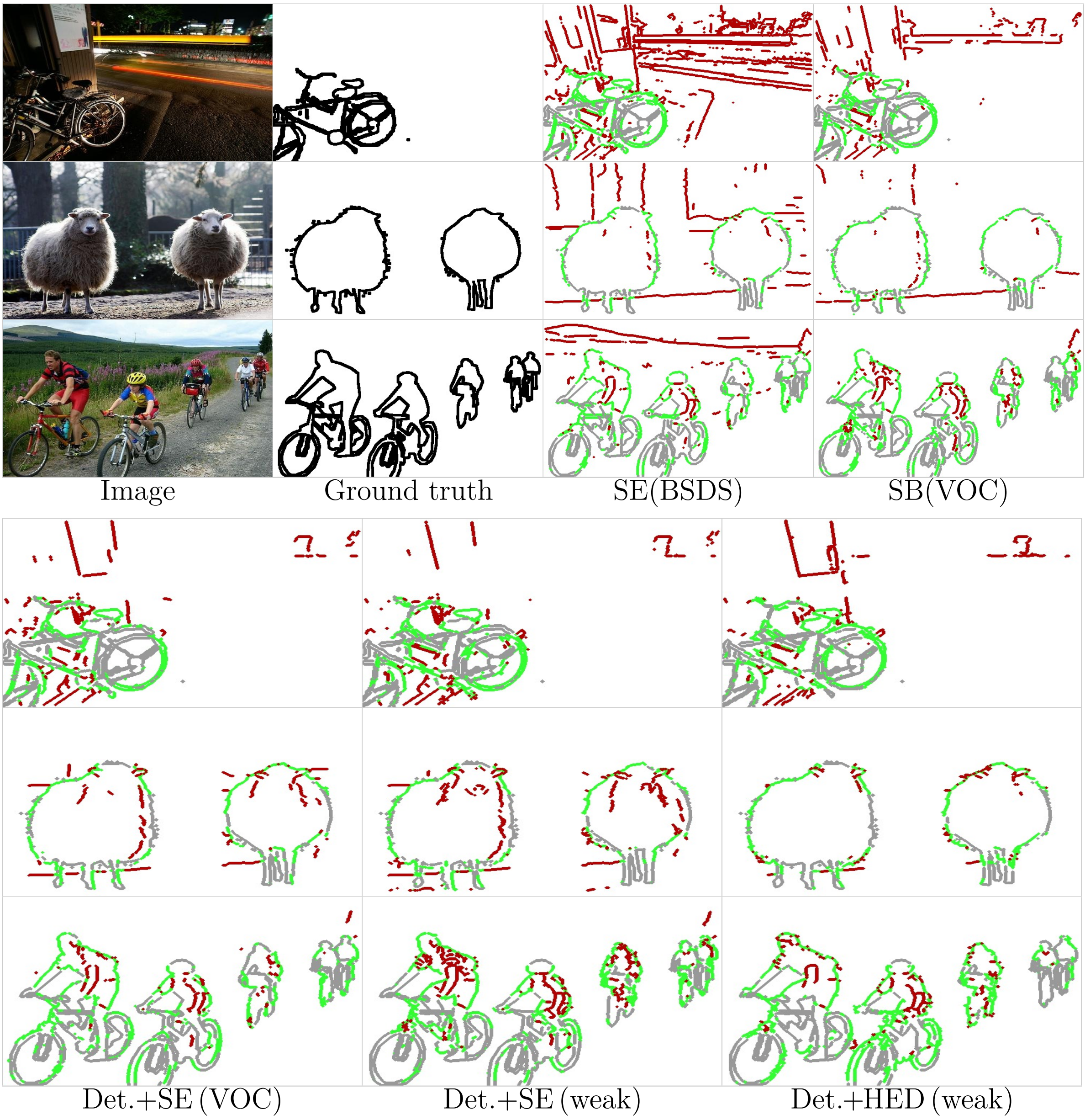


Figure 8.6: Qualitative results on *VOC*. $(\cdot)$ denotes the data used for training. Red/green indicate false/true positive pixels, grey is missing recall. All methods are shown at 50% recall. *Det.+SE (weak)* refers to the model *Det.+SE* $(SeSe_+ \cap BBs)$ *Det.+HED (weak)* refers to *Det.+HED* $(cons.S\&G \cap BBs)$. Object-specific boundaries differ from generic boundaries (such as the ones detected by *SE(BSDS)*). By using an object detector we can suppress non-object boundaries and focus boundary detection on the classes of interest. The proposed weakly supervised techniques allow to achieve high quality boundary estimates that are similar to the ones obtained by strongly supervised methods.

	Family	Method	mF	mAP
Other	GT	Inverse Detectors (Hariharan *et al.*, 2011)	28	21
SE	GT	SB(SBD) orig. (Uijlings and Ferrari, 2015)	39	32
		SB(SBD)	43	37
		Det.+SE (SBD)	$\underline{51}$	$\underline{45}$
	Other GT	Det.+SE (BSDS)	51	44
		Det.+MCG (BSDS)	50	42
	Weakly super-vised	SB(SeSe ∩ BBs)	40	34
		SB (MCG ∩ BBs)	42	35
		Det.+SE (SeSe ∩ BBs)	48	42
		Det.+SE (MCG ∩ BBs)	**51**	**45**
HED	GT	HED (SBD)	44	41
		Det.+HED (SBD)	$\underline{49}$	$\underline{45}$
	Other GT	HED(BSDS)	38	32
		Det.+HED (BSDS)	49	44
	Weakly super-vised	HED(cons. MCG ∩ BBs)	41	37
		HED (cons. S&G ∩ BBs)	44	39
		Det.+HED (cons. MCG ∩ BBs)	48	44
		Det.+HED (cons. S&G ∩ BBs)	**52**	**47**

Table 8.5: *SBD* results. Results are mean F(ODS)/AP across all 20 categories. (·) denotes the data used for training. See also Fig. 8.7. Underlined results correspond to baselines that rely on ground truth boundaries, and our best weakly supervised results are in boldface.

The results are summarised in Tab. 8.4. On the *COCO* benchmark for both *SE* and *HED* the models trained on the proposed weak annotations perform as well as the strongly supervised *SE* models. Similar to the *VOC* benchmark the *HED* model trained on ground truth shows superior performance.

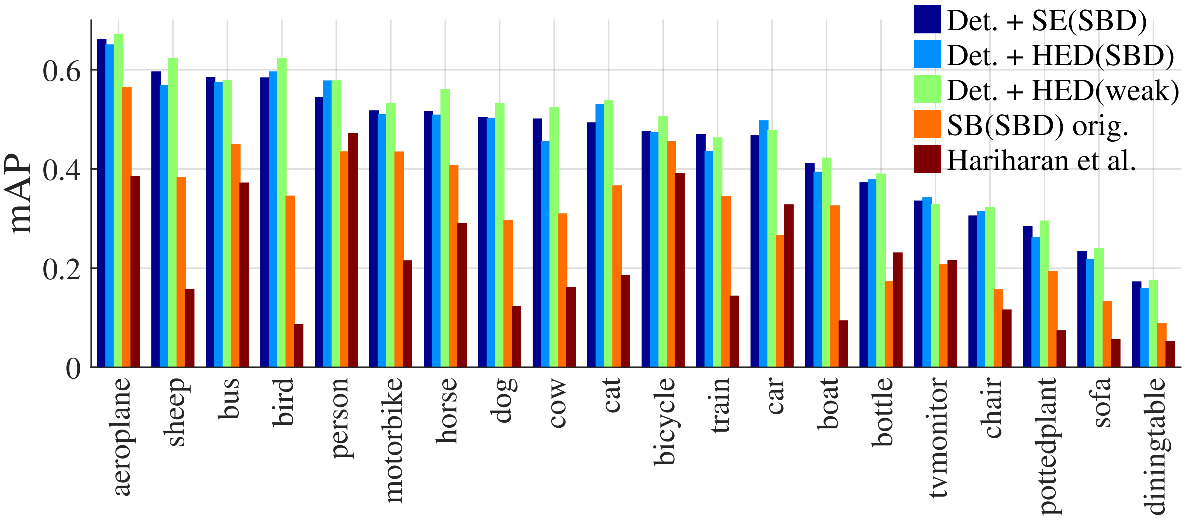


Figure 8.7: SBD results per class. (·) denotes the data used for training. Det.+HED (weak) refers to the model Det.+HED (cons. S&G ∩ BBs).

8.7 Results: Semantic Boundary Detection

In this section we analyse the performance of the proposed weakly supervised boundary variants trained with *SE* and *HED* on the *SBD* dataset (Hariharan *et al.*, 2011). In contrast to the *VOC* benchmark we move from object boundaries to class specific object boundaries. We are interested in external boundaries of all annotated objects of the specific semantic class and all internal boundaries are ignored during evaluation following the benchmark Hariharan *et al.* (2011). The results are presented in Fig. 8.7 and in Tab. 8.5.

Strongly supervised Applying the *SE* model plus object detection at test time outperforms the class-specific situational boundary detector (for both the original *SB* (Uijlings and Ferrari, 2015) and our re-implementation) as well as the *Inverse Detectors* method (Hariharan *et al.*, 2011). The model trained with SE on ground truth performs as well as the *HED* detector. Both of the models are good at detecting external object boundaries, however *SE* as it considers more local inputs triggers more on internal boundaries than *HED*. In the *VOC* evaluation detecting internal object boundaries is penalised, while in *SBD* these are ignored. This explains the small gap in the performance between *SE* and *HED* on this benchmark.

Weakly supervised The models trained with the proposed weakly-supervised boundary variants perform on par with the strongly supervised detectors, while only using bounding boxes or generic boundary annotations. We show in Tab. 8.5 the top result with the *Det. + HED(cons. S&G∩BBs)* model, achieving the state-of-the-art performance on the *SBD* benchmark. As Fig. 8.7 shows our weakly supervised approach considerably outperforms *SB* (Uijlings and Ferrari, 2015) and *Inverse Detectors* (Hariharan *et al.*, 2011) on all 20 classes.

8.8 Conclusion

In this chapter, we presented experiments which demonstrate that high quality object boundaries can be detecting using bounding box annotations. Relying on these alone, our proposed weakly-supervised training already improves over previously reported strongly supervised results for object-specific boundaries. When using generic boundary or ground truth annotations, we achieve the top performance on the object boundary detection task at the time, outperforming previously reported results by a large margin.

Part III

SHAPE AND POSE RECOVERY

Neural Body Fitting: 3D Human Shape and Pose Recovery

In previous chapters, we considered the tasks of person localisation and pixel-wise prediction. Here, we go one step further and address the task of predicting 3D human body pose and shape.

This is a challenging task even for highly parametrised deep learning models. Mapping from the 2D image space to the prediction space is difficult: perspective ambiguities make the loss function noisy and training data is scarce.

We tackle this problem with a novel approach we call *Neural Body Fitting* (NBF) that marries aspects of direct prediction and model-based approaches. This involves incorporating a model of the human body into a deep learning architecture, which has several advantages. First, the model incorporates limb orientations and shape, which are required for many applications such as character animation, biomechanics and virtual reality. Second, anthropomorphic constraints are automatically satisfied — for example limb proportions and symmetry. Third, the 3D model output is one step closer to a faithful 3D reconstruction of people in images.

In detailed experiments, we analyse how the components of our model affect performance, especially the use of part segmentations as an explicit intermediate representation prior to lifting, and present a robust, efficiently trainable framework for 3D human pose estimation from 2D images with competitive results on standard benchmarks.

This work was published at 3DV (Omran *et al.*, 2018) and won the best student paper award. All the experiments and analysis were conducted by Mohamed Omran, and Christoph Lassner provided help with the implementation. Gerard Pons-Moll and Peter Gehler contributed to the writing and discussion[9].

9.1 Introduction

Our goal is to fit an articulated 3D mesh of a human to a single monocular image (Fig. 9.1), thus recovering both body shape and pose. Traditional *model-based* approaches typically optimise an objective function that measures how well a body model fits the image observations — for example, 2D keypoints (Bogo *et al.*, 2016; Lassner *et al.*, 2017). These methods do not require paired 3D training data (images with 3D pose),

[9] We would also like to thank Dingfan Chen for help with re-training *HMR* Kanazawa *et al.* (2018).

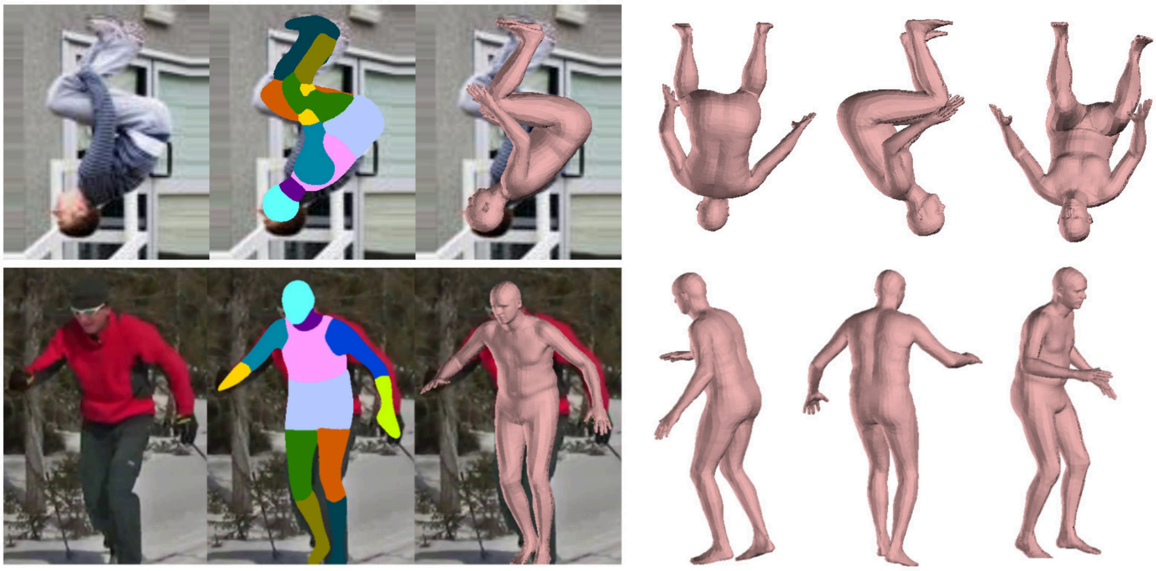


Figure 9.1: Given a single 2D image of a person, our goal is to recover a rich 3D reconstruction of the body. As a first step, we predict a semantic body part segmentation. This is provided in colour-coded form to a lifting network which predicts the parameters of a 3D body model.

but only work well when initialised close to the solution. By contrast, initialisation is not required in *learning-based* approaches, such as those based on convolutional neural networks (CNNs), which directly predict the desired 3D output. However these methods typically require many images with 3D shape and pose annotations, which are difficult to obtain unlike images with 2D keypoint annotations.

We therefore propose a hybrid architecture — *Neural Body Fitting* (*NBF*) — that integrates a statistical body model within a CNN, allowing us to directly predict shape and pose while taking top-down body model constraints into account. Specifically, from an image, a CNN predicts the parameters of the *SMPL* body model (Loper *et al.*, 2015), and the model is re-projected onto the image to evaluate the loss function in 2D space. Consequently, 2D keypoint annotations can be used to train such architectures reducing the need for 3D annotations. A few recent works have proposed very similar architectures that are trained using model-based loss functions (Tung *et al.*, 2017a; Kanazawa *et al.*, 2018; Pavlakos *et al.*, 2018b). While all these hybrid approaches share similarities, they all differ in essential design choices, such as the amount of 3D vs. 2D annotations for supervision and the input representation used to lift to 3D. (See Sec. 3.3.3 for an extensive discussion of these and other subsequent methods that follow the same approach.)

One key question we address with our study is whether to use an intermediate representation rather than directly lifting to 3D from the raw RGB image. Images of humans can vary due to factors such as illumination, clothing, and background clutter. Those effects do not necessarily correlate with pose and shape, thus we investigate

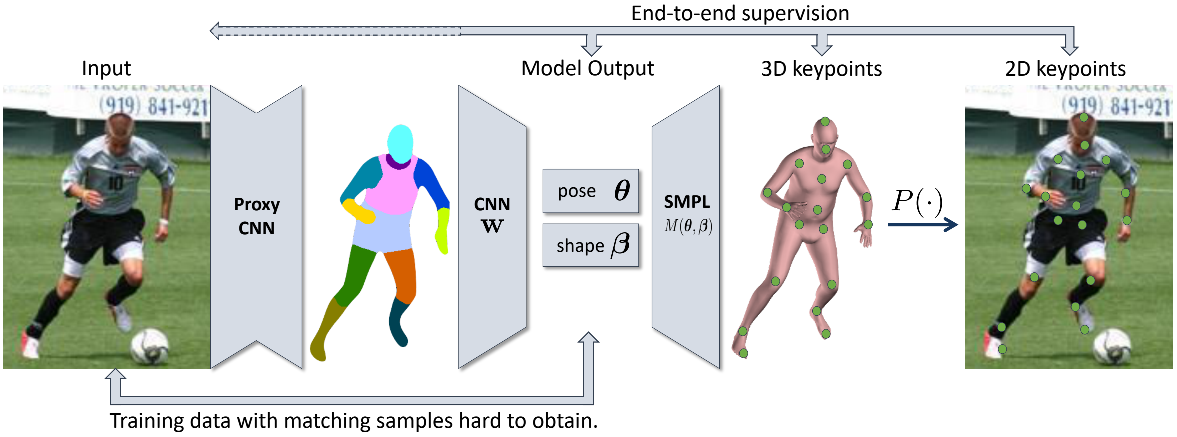


Figure 9.2: *Summary of our proposed pipeline.* We process the image with a standard semantic segmentation CNN into 12 semantic parts (see Sec. 9.3.2). An encoding CNN processes the semantic part probability maps to predict *SMPL* body model parameters (see Sec. 9.2.2), then via an embedded *SMPL* model produces a projection of the pose-defining keypoints to 2D. With these keypoints, a loss on 2D vertex positions can be backpropagated through the entire model (see Sec. 9.2.3).

whether a simplification of the RGB image into a semantic segmentation of body parts improves 3D inference. We also consider the granularity of the body part segmentation as well as segmentation quality, and find that:

- a colour-coded 12-body-part segmentation contains sufficient information for reliably predicting shape and pose,
- the use of such an intermediate representation results in competitive performance and easier, more data-efficient training compared to similar methods that predict pose and shape parameters from raw RGB images,
- segmentation quality is a strong predictor of shape and pose fit quality.

We also demonstrate that only a small fraction of the training data needs to be paired with 3D annotations. We make use of the *UP-3D* dataset (Lassner *et al.*, 2017) that consists of 8515 images in the wild along with 3D pose annotations. Larger 2D datasets exist, but *UP-3D* allows us to perform a controlled study.

9.2 Method

There are two main stages in the proposed architecture (see Fig. 9.2 for an overview): In the first stage, a body part segmentation is predicted from the RGB image. The second stage takes this segmentation to predict the body model parameters: a low-dimensional parametrisation of a mesh. Those parameters are passed to *SMPL* (Loper *et al.*, 2015) to produce a 3D mesh attached to a skeleton. The skeleton joints are then projected to the image closing the loop. Hence, *NBF* admits both full 3D supervision (in the model or 3D Euclidean space) and weak 2D supervision (if images with only 2D annotations are available).

9.2.1 Body Model

For our experiments we use the *SMPL* body model due to its good trade-off between high anatomic flexibility and realism. *SMPL* parametrises a triangulated mesh with $N = 6890$ vertices with pose parameters $\theta \in \mathbb{R}^{72}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ – optionally the translation parameters $\gamma \in \mathbb{R}^3$ can be taken into account as well.

Shape $B_s(\boldsymbol{\beta})$ and pose dependent deformations $B_p(\boldsymbol{\theta})$ are first applied to a base template $\mathbf{T}_\mu$; then the mesh is posed by rotating each body part around skeleton joints $J(\boldsymbol{\beta})$ using a skinning function W:

$$\mathit{SMPL}\ (\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}), \tag{9.1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}), \tag{9.2}$$

where $\mathit{SMPL}\ (\boldsymbol{\beta}, \boldsymbol{\theta})$ is the *SMPL* function, and $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ outputs an intermediate mesh in a T-pose after pose and shape deformations are applied. *SMPL* produces realistic results using relatively simple mathematical operations – most importantly for us *SMPL* is fully differentiable with respect to pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$. All these operations, including the ones to determine projected points of a posed and parametrised 3D body can be represented as layers of a neural network. We use them to make the 3D body a part of our deep learning model.

9.2.2 Lifting Network

NBF predicts the parameters of the body model from a colour-coded part segmentation map $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$ using a CNN-based predictor parametrised by weights w. The estimators for pose and shape are thus given by $\boldsymbol{\theta}(w, \mathbf{I})$ and $\boldsymbol{\beta}(w, \mathbf{I})$ respectively.

We integrate the *SMPL* model and a simple 2D projection layer into our CNN estimator, as described in Sec. 9.2.1. This allows us to output a 3D mesh, 3D skeleton joint locations or 2D joints, depending on the kind of supervision we want to apply for training while keeping the CNN monolithic.

Mathematically, the function $N_{3D}(w, \mathbf{I})$ that maps from semantic images to meshes is given by

$$\begin{aligned} N_{3D}(w, \mathbf{I}) &= M(\boldsymbol{\theta}(w, \mathbf{I}), \boldsymbol{\beta}(w, \mathbf{I})) && (9.3) \\ &= W(T(\boldsymbol{\beta}(w, \mathbf{I}), \boldsymbol{\theta}(w, \mathbf{I}), \\ &\quad J(\boldsymbol{\beta}(w, \mathbf{I})), \boldsymbol{\theta}(w, \mathbf{I}), \mathbf{W})), && (9.4) \end{aligned}$$

which is the *SMPL* equation (Eq. (9.1)) parametrised by network weights w. From this it is obvious that we can easily find the derivatives $\frac{\partial N_{3D}}{\partial w}$ by using chain rule. *NBF* can also predict the 3D joints $N_J(w, \mathbf{I}) = J(\boldsymbol{\beta}(w, \mathbf{I})$, because they are a function of the model parameters. Furthermore, using a projection operation $\pi(\cdot)$ we can project the 3D joints onto the image plane

$$N_{2D}(w, \mathbf{I}) = \pi(J(w, \mathbf{I})), \qquad (9.5)$$

where $N_{2D}(w, \mathbf{I})$ is the *NBF* function that outputs 2D joint locations. All of these operations are differentiable and allow us to use gradient-based optimisation to update model parameters with a suitable loss function.

9.2.3 Loss Functions

We experiment with the following loss functions:

3D latent parameter loss: This is an L1 loss on the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Given a paired dataset $\{\mathbf{I}_i, \boldsymbol{\theta}_i, \boldsymbol{\beta}_i\}_i^N$, the loss is given by:

$$\mathcal{L}_{lat}(w) = \sum_i^N |\mathbf{r}(\boldsymbol{\theta}(w, \mathbf{I}_i)) - \mathbf{r}(\boldsymbol{\theta}_i)| + |\boldsymbol{\beta}(w, \mathbf{I}_i) - \boldsymbol{\beta}_i|, \qquad (9.6)$$

where $\mathbf{r}$ are the vectorised rotation matrices of the 23 parts of the body together with a global rotation matrix. Similar to Lassner *et al.* (2017); Pavlakos *et al.* (2018b), we observed better performance by imposing the loss on the rotation matrix representation of $\boldsymbol{\theta}$ rather than on its 'native' axis angle encoding as defined in *SMPL* . This requires us to project the predicted matrices to the manifold of rotation matrices. We perform this step using singular value decomposition (SVD) to maintain differentiability.

3D joint loss: Given a paired dataset with skeleton annotations $\{\mathbf{I}_i, \boldsymbol{\theta}_i, \mathbf{J}\}_i^N$ we compute the loss in terms of 3D joint position differences as:

$$\mathcal{L}_{3D}(w) = \sum_i^N \|N_J(w, \mathbf{I}_i) - \mathbf{J}_i\|^2 \qquad (9.7)$$

2D joint loss: If the dataset $\{\mathbf{I}_i, \mathbf{J}_{2D}\}_i^N$ provides solely 2D joint position ground truth, we define a similar loss in terms of 2D distance and rely on error backpropagation through the projection:

$$\mathcal{L}_{2D}(w) = \sum_i^N \|N_{2D}(w, \mathbf{I}_i) - \mathbf{J}_{2D,i}\|^2 \qquad (9.8)$$

Joint 2D and 3D loss: To maximise the amounts of usable training data, ideally multiple data sources can be combined with a subset of the data $\mathcal{D}_{3D}$ providing 3D annotations and another subset $\mathcal{D}_{3D}$ providing 2D annotations. We can trivially integrate all the data with different kinds of supervision by falling back to the relevant losses and setting them to zero if not applicable.

$$\mathcal{L}_{2D+3D}(w, \mathcal{D}) = \mathcal{L}_{2D}(w, \mathcal{D}_{2D}) + \mathcal{L}_{3D}(w, \mathcal{D}_{3D}) \tag{9.9}$$

In our experiments, we analyse the performance of each loss and their combinations. In this work, we are mostly interested in the last case scenario, with a mixture of 3D and 2D data. In particular, we evaluate how much gain in 3D estimation accuracy can be obtained from weak 2D annotations which are much cheaper to obtain than accurate 3D annotations.

9.3 Results

9.3.1 Experimental Settings

We used the following three datasets for evaluation: *UP-3D* (Lassner *et al.*, 2017), *HumanEva-I* (Sigal *et al.*, 2010), and *Human3.6M* (*H36M*) (Ionescu *et al.*, 2014). These datasets are described in detail in Chapter 3.

We perform a detailed analysis of our approach on *UP-3D* and *Human3.6M*, and compare against state-of-the-art methods on *HumanEVA-I* and *Human3.6M*. Sec. 3.2 contains detailed descriptions of these datasets. For our analysis on *UP-3D*, we use the training (5703 images) and validation (1423 images) sets. For experiments on *H36M*, we reserve subjects S1, S5, S6 and S7 for training, and hold out subject S8 for validation. We compare to the state of art on the test sequences S9 and S11.

9.3.2 Implementation

Data preparation To train our model, we require images paired with 3D body model fits (i.e. *SMPL* parameters) as well as pixel-wise part labels. The UP-3D dataset provides such annotations, while *Human3.6M* does not. However, by applying *MoSH* (Loper *et al.*, 2014) to the 3D mocap marker data provided by the latter we obtain the corresponding *SMPL* parameters, which in turn allows us to generate part labels by rendering an appropriately annotated *SMPL* mesh (Lassner *et al.*, 2017).

Scale ambiguity The *SMPL* shape parameters encode among other factors a person's size. Additionally, both distance to the camera and focal length determine how large a person appears in an image. To eliminate this ambiguity during training, we constrain scale information to the shape parameters by making the following assumptions: The camera is always at the *SMPL* coordinate origin, the optical axis always points in the same direction, and a person is always at a fixed distance from the camera. We render the ground truth *SMPL* fits and scale the training images to fit the renderings (using the corresponding 2D joints). This guarantees that the the only factor affecting person size in the image are the *SMPL* shape parameters. At test-time, we estimate person height and location in the image using 2D *DeeperCut* keypoints (Insafutdinov *et al.*, 2016), and centre the person within a crop of 512 × 512 pixels (px) such that they have a height of $440px$, which roughly corresponds to the setting seen during training.

Architecture We use a two-stage approach: The first stage receives the 512px × 512px input crop and produces a part segmentation. We use our own re-implementation of the *RefineNet* (Lin *et al.*, 2017a) semantic segmentation network, which uses *ResNet-101* (He *et al.*, 2016)) as a feature extraction backbone. The resulting part segmentation is then colour-coded, resized to 224px × 224px and fed as an RGB image to the second stage. The latter is based on a repurposed *ResNet-50*) network. We replace the final pooling layer with a single fully-connected layer that outputs the 10 shape and 216 pose parameters of *SMPL* . This is followed by a non-trainable set of layers that implement the *SMPL* model and an image projection. Such layers can produce a 3D mesh, 3D joints or 2D joints given the predicted pose and shape. We implement our method in *TensorFlow* (Abadi *et al.*, 2016)..

Training We train the segmentation network for 20 epochs with a batch size of 5 using the *ADAM* optimiser (Kingma and Ba, 2015). Learning rate and weight decay are set to 0.00002 and 0.0001 respectively, with a polynomial learning rate decay. For training the segmentation network on *UP-3D* we used the 5703 training images. For *Human3.6M* we subsampled the videos, only using every 10th frame from each video, which results in about 32000 frames. Depending on the amount of data, training the segmentation networks takes about 6-12 hours on a Volta V100 machine. We train the fitting network for 75 epochs with a batch size of 5 also using *ADAM*. The learning rate is set to 0.00004 with polynomial decay and we use a weight decay setting of 0.0001. We found that an L1 loss on the *SMPL* parameters was a little better than an L2 loss. We also experimented with robust losses (e.g. Geman-McClure (Geman and McClure, 1987) and Tukey's biweight loss (Belagiannis *et al.*, 2015)) but did not observe benefits. Training this network takes about 1.5 hours for the *UP-3D* dataset and six hours for *Human3.6M*. Thus all in all training both stages requires a maximum total of 18 (12+6) hours on a single machine.

Data Augmentation At test-time we cannot guarantee that the person will be perfectly centred in the input crop, which can lead to degraded performance. We found it thus critical to train both the segmentation network and the fitting network with strong data augmentation, especially including random jitter, scaling ($0.9 - 1.1\times$), horizontal reflection (which requires re-mapping the labels), as well as rotations (up to 45 degrees).

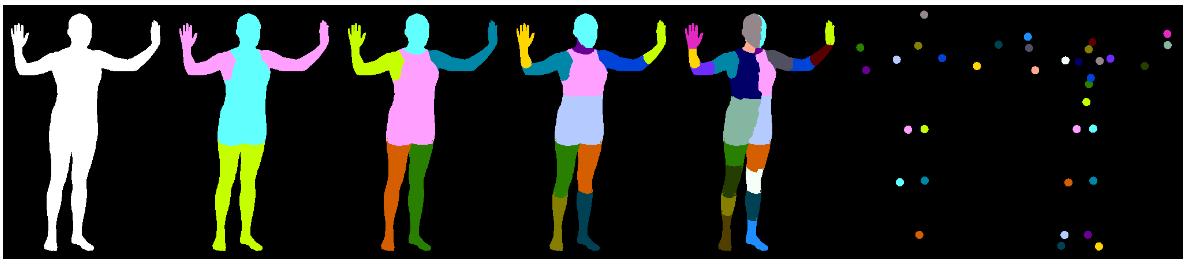


Figure 9.3: Example training image annotations illustrating different types of inputs at different levels of granularities. We generate these automatically from the corresponding 3D ground truth if available. From left to right: 1-, 3-, 6-, 12-, and 24-part segmentations, followed by 14- and 24-joint skeletons.

type of input	*UP*	*H36M*
RGB	98.5	48.9
Segmentation (1 part)	95.5	43.0
Segmentation (3 parts)	36.5	37.5
Segmentation (6 parts)	29.4	36.2
Segmentation (12 parts)	27.8	33.5
Segmentation (24 parts)	28.8	31.8
Joints (14)	28.8	33.4
Joints (24)	27.7	33.4

Table 9.1: Input Type vs. 3D error in millimeters

9.3.3 Analysis

Which Input Encoding? We investigate here what input representation is effective for pose and shape prediction. Full RGB images certainly contain more information than for example silhouettes, part segmentations or 2D joints. However, some information may not be relevant for 3D inference, such as appearance, illumination or clothing, which might make the network overfit to nuisance factors

To this end, we train a network on different image representations and compare their performance on the *UP-3D* and *Human3.6M* validation sets. We compare RGB images, colour-coded part segmentations of varying granularities, and colour-coded joint heatmaps (see Sec. 9.3.3 for examples). We generate both using the ground truth *SMPL* annotations to establish an upper bound on performance, and later consider the case where we do not have access to such information at test time.

The results are reported in Tab. 9.1. We observe that explicit part representations (part segmentations or joint heatmaps) are more useful for 3D shape/pose estimation compared to RGB images and plain silhouettes. The difference is especially pronounced on the *UP-3D* dataset, which contains more visual variety than the images of *Human3.6M*,

Val \ Train	VGG	ResNet	RefineNet	GT
VGG-16	107.2	119.9	135.5	140.7
ResNet	97.1	96.3	112.2	115.6
RefineNet	89.6	89.9	82.0	83.3
GT	62.3	60.5	35.7	27.8

Table 9.2: *Effect of segmentation quality on the quality of the 3D fit prediction modules ($err_{joints3D}$)*

with an error drop from 98.5 mm to 27.8 mm when using a 12 part segmentation. This demonstrates that a 2D segmentation of the person into sufficient parts carries a lot of information about 3D pose/shape, while also providing full spatial coverage of the person (compared to joint heatmaps). Is it then worth learning separate mappings first from image to part segmentation, and then from part segmentation to 3D shape/pose? To answer this question we first need to examine how 3D accuracy is affected by the quality of real predicted part segmentations.

Which Input Quality? To determine the effect of segmentation quality on the results, we train three different *part segmentation networks*. Besides *RefineNet*, we also train two variants of *DeepLab* (Chen *et al.*, 2015a), based on *VGG-16* (Simonyan and Zisserman, 2015) and *ResNet-101* (He *et al.*, 2016). These networks result in IoU scores of 67.1, 57.0, and 53.2 respectively on the *UP-3D* validation set. Given these results, we then train four *3D prediction networks* — one for each of the part segmentation networks, and an additional one using the ground truth segmentations. We report 3D accuracy on the validation set of *UP-3D* for each of the four 3D networks, diagonal numbers of Tab. 9.2. As one would expect, the better the segmentation, the better the 3D prediction accuracy. As can also be seen in Tab. 9.2, better segmenters at test time always lead to improved 3D accuracy, even when the 3D prediction networks are trained with poorer segmenters. This is perhaps surprising, and it indicates that mimicking the statistics of a particular segmentation method at training time plays only a minor role. For example a network trained with GT segmentations and tested using *RefineNet* segmentations performs on par with a network that is trained using *RefineNet* segmentations (83.3mm vs 82mm). To further analyse the correlation between segmentation quality and 3D accuracy, in Fig. 9.4 we plot the relationship between F1-score and 3D reconstruction error. Each dot represents one image, and the colour its respective difficulty — we use the distance to mean pose as a proxy measure for difficulty. The plot clearly shows that the higher the F1-score, the lower the 3D joint error.

Which Types of Supervision? We now examine different combinations of loss terms. The losses we consider are L_{lat} (on the latent parameters), L_{3D} (on 3D joint/vertex locations), L_{2D} (on the projected joint/vertex locations). We compare performance using three different error measures: (i) $err_{joints3D}$, the Euclidean distance between ground truth and predicted *SMPL* joints (in mm). (ii) *PCKh* (Andriluka *et al.*, 2014),

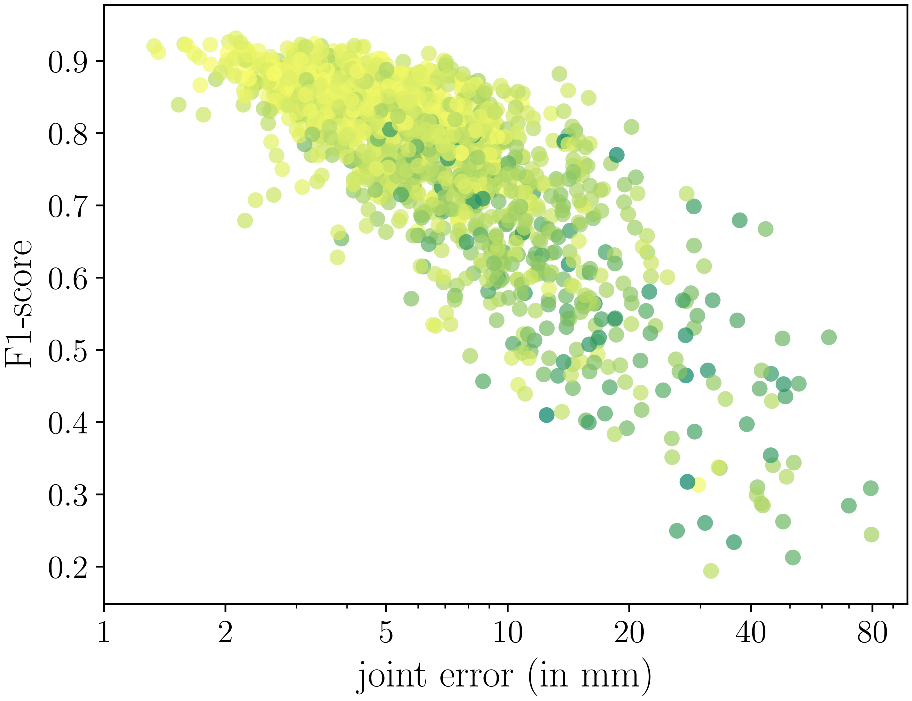


Figure 9.4: *Segmentation quality (F1-score) vs. fit quality (3D joint error).* The darkness indicates the difficulty of the pose, i.e. the distance from the upright pose with arms by the sides.

the percentage of correct keypoints with the error threshold being 50% of head size, which we measure on a per-example basis. (iii) err_{quat}, quaternion distance error of the predicted joint rotations (in radians).

Given sufficient data — the full 3D-annotated *UP-3D* training set with mirrored examples (11406) — only applying a loss on the model parameters yields reasonable results, and in this setting, additional loss terms don't provide benefits. When only training with L_{3D}, we obtain similar results in terms of $err_{joints3D}$, however, interestingly err_{quat} is significantly higher. This indicates that predictions produce accurate 3D joints positions in space, but the limb orientations are incorrect. This further demonstrates that methods trained to produce only 3D keypoints do not capture orientation, which is needed for many applications.

We also observe that only training with the 2D reprojection loss (perhaps unsurprisingly) results in poor performance in terms of 3D error, showing that some amount of 3D annotations are necessary to overcome the ambiguity inherent to 2D keypoints as a source of supervision for 3D.

Due to the *SMPL* layers, we can supervise learning with any number of joints/mesh vertices. We thus experimented with the 91 landmarks used by Lassner *et al.* (2017) for their fitting method but find that the 24 *SMPL* joints are sufficient in this setting.

Loss	$\mathbf{err_{joints3D}}$	**PCKh**	$\mathbf{err_{quat}}$
L_{lat}	83.7	93.1	0.278
$L_{lat} + L_{3D}$	82.3	93.4	0.280
$L_{lat} + L_{2D}$	83.1	93.5	0.278
$L_{lat} + L_{3D} + L_{2D}$	82.0	93.5	0.279
L_{3D}	83.7	93.5	1.962
L_{2D}	198.0	94.0	1.971

Table 9.3: *Loss ablation study.* Results in 2D and 3D error metrics (*joints3D*: Euclidean 3D distance, *mesh*: average vertex to vertex distance, *quat*: average body part rotation error in radians).

Ann.perc. Error	100	50	20	10	5	2	1	0
$\mathbf{err_{joints3D}}$	83.1	82.8	82.8	83.6	84.5	88.1	93.9	198
$\mathbf{err_{quat}}$	0.28	0.28	0.27	0.28	0.29	0.30	0.33	1.97

Table 9.4: *Effect of 3D labelled data.* We show the 3D as well as the estimated body part rotation error for varying ratios of data with 3D labels. For all of the data, we assume that 2D pose labels are available. Both errors saturate at 20% of 3D labelled training examples.

How Much 3D Supervision Do We Need? The use of these additional loss terms also allows us to leverage data for which no 3D annotations are available. With the following set of experiments, we attempt to answer two questions: (i) Given a small amount of 3D-annotated data, does extra 2D-annotated data help?, (ii) What amount of 3D data is necessary? To this end we train multiple networks, each time progressively disabling the 3D latent loss and replacing it with the 2D loss for more training examples. The results are depicted in Sec. 9.3.3. We find that performance barely degrades as long as we have a small amount of 3D annotations. In contrast, using small amounts of 3D data and no extra data with 2D annotations yields poor performance. This is an important finding since obtaining 3D annotations is difficult compared to simple 2D keypoint annotations.

Qualitative Results A selection of qualitative results from the *UP-3D* dataset can be found in Sec. 9.3.3. We show examples from the four different error quartiles. Fits from the first three quartiles still reproduce the body pose somewhat faithfully, and only in the last row and percentile, problems become clearly visible. To illustrate the high correlation between input segmentation quality and output fit quality, we present the four worst examples from the validation set in terms of 3D joints reconstruction error when (i) we use our trained part segmentation network (Fig. 9.6a), and when (ii) the network is trained to predict body model parameters from ground truth segmentations (Fig. 9.6b). In the latter case, there are still errors but these are noticeably less severe.

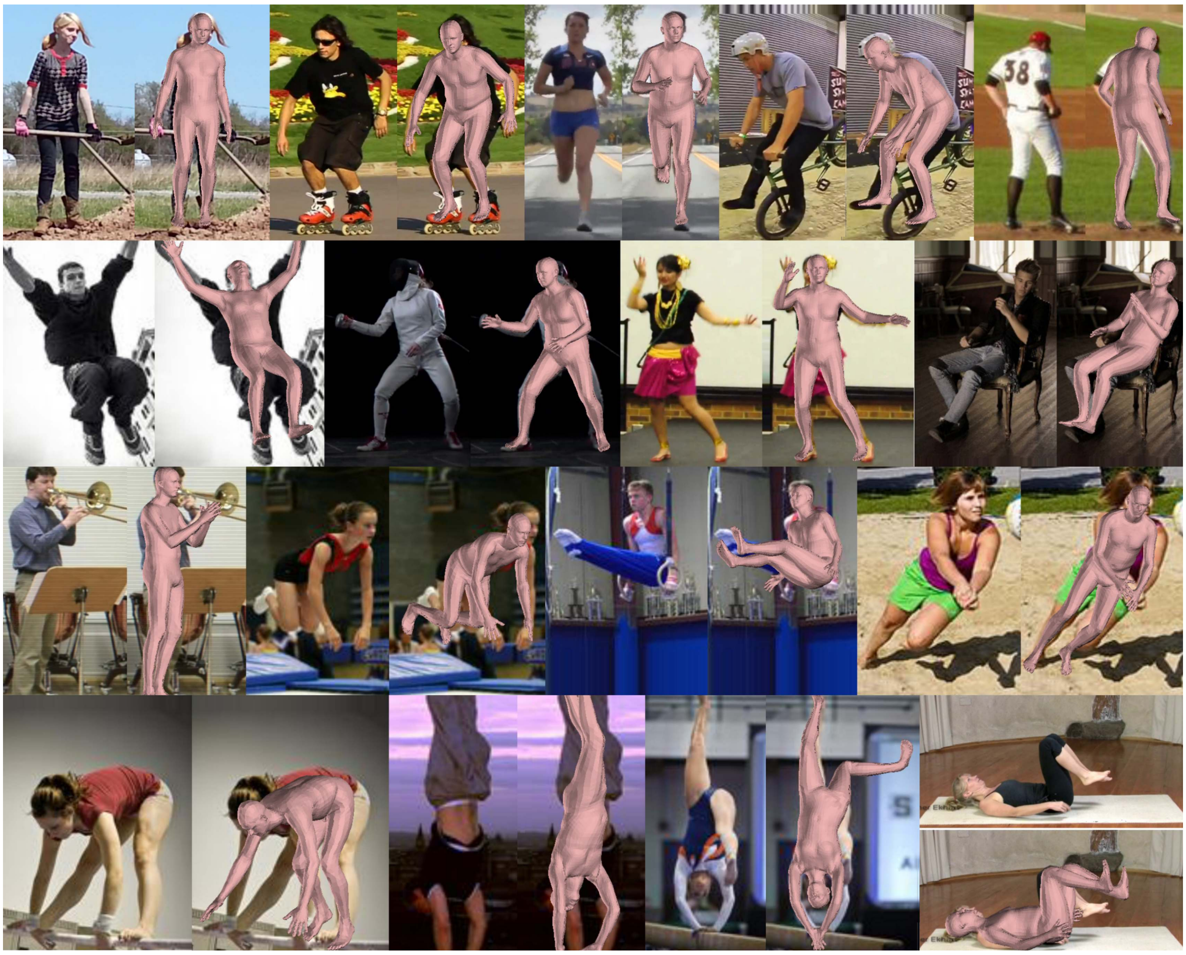


Figure 9.5: *Qualitative results by error quartile in terms of* $err_{joints3D}$. The rows show representative examples from different error quartiles, top to bottom: 0-25%, 25-50%, 50-75%, 75-100%

9.3.4 Comparison to State-of-the-Art

Here we compare to the state of the art on *HumanEva-I* (Tab. 9.5) and *Human3.6M* (Tab. 9.6). We perform a per-frame rigid alignment of the 3D estimates to the ground truth using Procrustes Analysis and report results in terms of reconstruction error, i.e. the mean per joint position error after alignment (given in *mm*). The model we use here is trained on *Human3.6M* data.

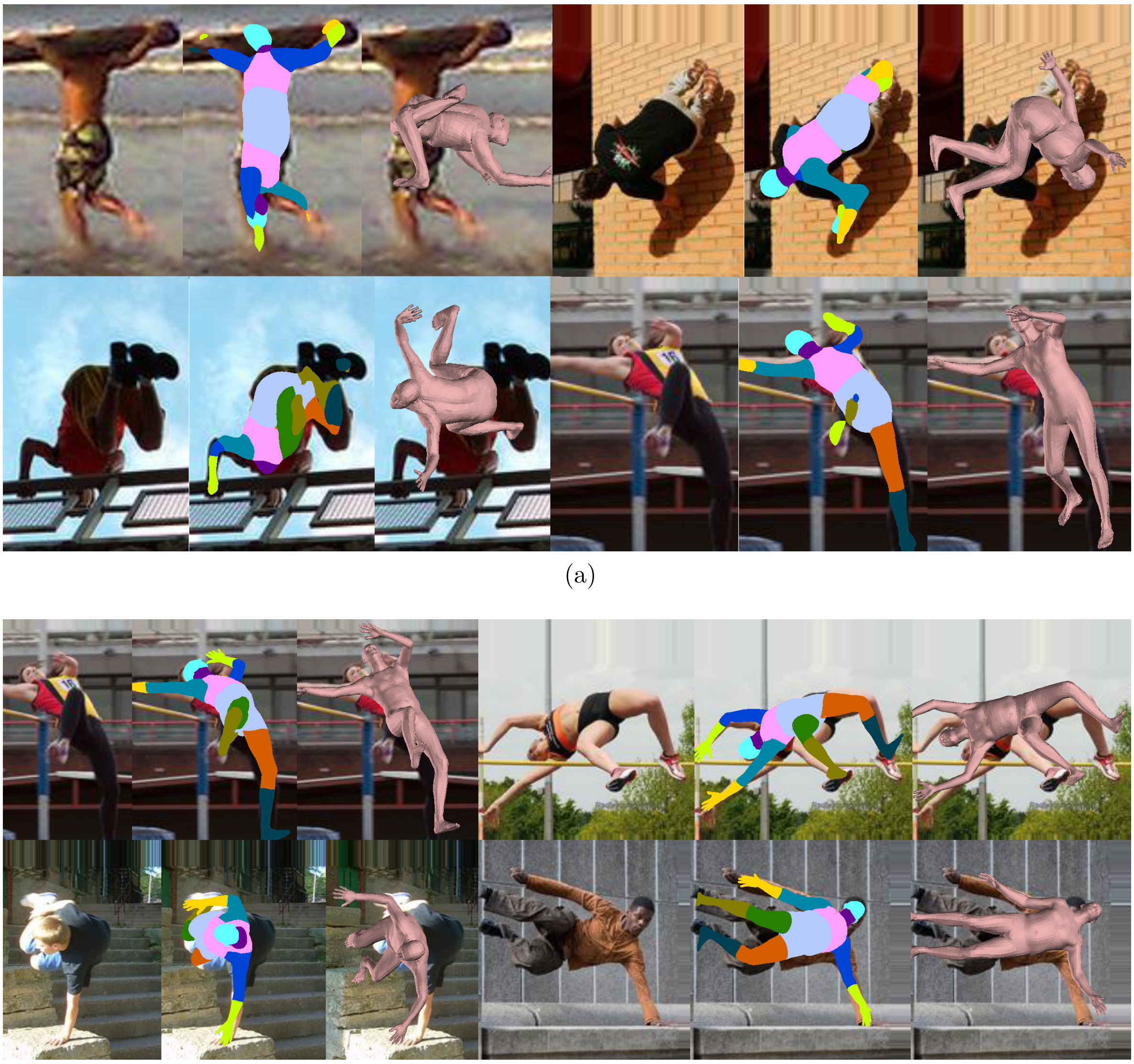
We compare favourably to similar methods, but these are not strictly comparable since they train on different datasets. Pavlakos *et al.* (2018b) do not use any data from *Human3.6M*, whereas *HMR* (Kanazawa *et al.*, 2018) does, along with several other datasets. We retrained the latter with the original code only using *Human3.6M* data for a more direct comparison to ours (*HMR* (*H36M*-trained) in Tab. 9.6). Given Tab. 9.1, we hypothesise that their approach requires more training data for good performance because it uses RGB images as input.

Method	Mean	Median
Ramakrishna et al. (Ramakrishna *et al.*, 2012)	168.4	145.9
Zhou et al. (Zhou *et al.*, 2015b)	110.0	98.9
SMPLify (Bogo *et al.*, 2016)	79.9	61.9
Random Forests (Lassner *et al.*, 2017)	93.5	77.6
SMPLify (Dense) (Lassner *et al.*, 2017)	74.5	59.6
Ours	64.0	49.4

Table 9.5: ***HumanEva-I* results.** 3D joint errors in mm.

Method	Mean	Median
Akhter & Black (Akhter and Black, 2015)	181.1	158.1
Ramakrishna et al. (Ramakrishna *et al.*, 2012)	157.3	136.8
Zhou et al. (Zhou *et al.*, 2015b)	106.7	90.0
SMPLify (Bogo *et al.*, 2016)	82.3	69.3
SMPLify (dense) (Lassner *et al.*, 2017)	80.7	70.0
SelfSup (Tung *et al.*, 2017a)	98.4	-
Pavlakos et al. (Pavlakos *et al.*, 2018b)	75.9	-
HMR (H36M-trained) (Kanazawa *et al.*, 2018)	77.6	72.1
HMR (Kanazawa *et al.*, 2018)	**56.8**	-
Ours	59.9	52.3

Table 9.6: ***Human3.6M.*** 3D joint errors in mm.



(a)

(b)

Figure 9.6: Worst examples from the validation set in terms of 3D error when the fitting network is provided with (a) imperfect segmentations, and (b) ground truth segmentations at test-time. In each case we train a separate fitting network on the corresponding inputs.

9.4 Conclusion

In this chapter, we make several principled steps towards a full integration of parametric 3D human pose models into deep CNN architectures. We analyse (1) how the 3D model can be integrated into a deep neural network, (2) how loss functions can be combined and (3) how a training can be set up that works most efficiently with scarce 3D data.

In contrast to existing methods we use a region-based 2D representation, namely a 12-body-part segmentation, as an intermediate step prior to the mapping to 3D shape and pose. This segmentation provides full spatial coverage of a person as opposed to the commonly used sparse set of keypoints, while also retaining enough information about the arrangement of parts to allow for effective lifting to 3D.

We used a stack of CNN layers on top of a segmentation model to predict an encoding in the space of 3D model parameters, followed by instantiation of an articulated body mesh and a projection of the corresponding skeleton to the image plane. This full integration allows us to finely tune the loss functions and enables end-to-end training. We found a loss that combines 2D as well as 3D information to work best. The flexible implementation allowed us to experiment with the 3D losses only for parts of the data, moving towards a weakly supervised training scenario that avoids expensive 3D labelled data. With 3D information for only 20% of our training data, we could reach similar performance as with full 3D annotations.

We believe that this encouraging result is an important finding for the design of future datasets and the development of 3D prediction methods that do not require expensive 3D annotations for training. Future work will involve extending this to more challenging settings involving multiple, possibly occluded, people. We also plan on exploring the use of part segmentations for supervision as well as test-time optimisation, as these provide complementary information to 2D keypoints, such as depth relations and self-occlusion.

10

Conclusions and Future Directions

In this thesis, we've addressed several different recognition tasks: pedestrian detection, different pixel labelling tasks — boundary detection, semantic segmentation, instance segmentation —, as well as 3D human shape and pose estimation. Before concluding with a discussion of possible future work, we will enumerate some conclusions.

Since the discussion of future work will partly touch upon aspects relevant to all of the problems we address, we need to first clarify the use of some terms. The tasks we addressed in this thesis and related ones such as image classification will be referred to as recognition tasks. We will refer to benchmarks commonly used to measure progress on these tasks (e.g. *Caltech*, *Cityscapes*, *MSCOCO*, *Human3.6M*, *3DPW*, *ImageNet*) as standard recognition benchmarks. When we refer to modern recognition methods, we mostly mean methods based on feedforward neural networks and CNNs in particular as these dominate the standard benchmarks at the time of writing. We distinguish between datasets and benchmarks, with the latter referring to the combination of a dataset and a suitable evaluation scheme.

10.1 Conclusions

10.1.1 Summary

Chapters 4 to 6 dealt with detection, specifically of pedestrians. In Chapter 4 we analysed a decade's worth of methods leading up to — and including the beginnings of — the deep learning boom. Among other things, this analysis demonstrated how critical image representations have been for performance. For the experiments we used a conceptually simple detector: a boosted decision forest consisting of level-2 trees operating on spatially-pooled input features. Our results showed that merely varying the input image representation was enough to replicate the leap in performance during the time period covered by our analysis. Starting with just intensity images, we matched the performance of the Viola-Jones detector Viola *et al.* (2003), and could continually improve results simply by adding further colour and gradient input feature channels. As a final step, we applied a small set of filters to these input feature channels (Nam *et al.*, 2014) which resulted in performance on par with the state of the art. Our analysis of the literature also showed the importance of context modelling and additional data modalities. Following these results, we augmented this simple detector with motion features and a simple person-to-person context model. This significantly outperformed the state-of-the-art.

In another set of experiments, we looked at cross-dataset generalisation. This involved training a detector on one dataset and evaluating it on others. Our results showed that detectors specialise on their respective training sets and do not generalise well to test sets from other sources. However, datasets vary in their suitability as generic training sets. In this particular setting, *INRIA* fared better than *Caltech* and *KITTI*, possibly due to the former being more diverse in terms of setting and of higher image quality.

Chapter 5 presented a simple CNN-based detector that did not rely on any problem-specific modelling. This outperformed competing methods that were explicitly designed to be sensitive to body parts and pedestrian-specific occlusion patterns. We showed that further gains were achievable by simply oversampling the training set as well as increasing the architecture size.

Chapter 6 concluded the section on detection with a forward-looking analysis complementary to the retrospective analysis of Chapter 4. There, we focused on the shortcomings of the then state of the art. We introduced a human baseline on the *Caltech* dataset that significantly outperformed pedestrian detectors across different evaluation settings. We then manually categorised errors into distinct groups: false positives (localisation, background, and annotation errors) as well as false negatives (small scale, occlusion, rare classes, and annotation errors). We then revisited the evaluation metric and find that it overlooks localisation errors — something that was foreshadowed by our results in Chapter 5. Additionally, we showed that cleaning up the annotations was very beneficial to CNN- and non-CNN-based methods. Both are sensitive to annotation noise, including imprecisely aligned bounding boxes. While CNNs provide stronger foreground-background disambiguation, the feature responses are somewhat diffuse, hence the need for additional bounding box regression for accurate localisation.

In Chapters 7 and 8, we focus on pixel-wise labelling tasks. Chapter 7 presents *Cityscapes*, a benchmark for two such tasks: pixel-level semantic labelling and instance-level semantic labelling. In the first, each pixel is assigned a semantic label and in the second, pixels additionally need to be grouped if they belong to a single instance. The latter is considered an extension of object detection that targets more precise outputs than a bounding box and is often addressed with similar methods. We present an empirical study together with the dataset which provides an in-depth analysis of its characteristics. We also evaluate several state-of-the-art approaches and demonstrate the dataset's difficulty. It has since established itself as a go-to benchmark for pixel-level and instance-level segmentation as well as for other uses, e.g. generative image modelling. A widely used pedestrian benchmark also relies on the data (Zhang *et al.*, 2017b).

Chapter 8 deals with the task of boundary detection. Given the difficulty of obtaining ground truth for this task, we explored the use of weak supervision with classical and CNN-based approaches. We derived pseudo-annotations for different variants of boundary detection, relying on classical unsupervised techiques, generic boundary detectors, and bounding box-based object detectors. We demonstrate that this can result in strong performance compared to both competing weakly- as well as fully-supervised methods.

Finally, we turn to the task of 3D human shape and pose estimation. We present a method that embeds a statistical body model in a neural network trained to regress shape and pose. This allows us to supervise the regressor with a mixture of data, both 3D and 2D. Unlike similar methods in the literature, we also use a part segmentation as an intermediate step. This provides an extra layer of interpretability and also results in data efficiency for the lifting step. We also showed a link between pose difficulty and rarity as well as segmentation quality and 3D pose accuracy.

10.2 Future Directions: An Overview

While each of the tasks we've addressed in this thesis has its own particularities, they naturally also have a lot in common. Our discussion of possible future directions will mostly by centred around these commonalities.

These tasks are related at the conceptual level in obvious ways: We seek to understand images of people at different levels of granularity, with the specific output being dictated by the task. All of these tasks also require some form of figure-ground organisation, i.e. separating an object from its background, either implicitly (Chapters 4 to 6), or as the explicit target output (Chapters 7 and 8), or as an explicit intermediate step (Chapter 9).

The methods we describe in this thesis also follow a common paradigm: supervised statistical learning in the i.i.d. setting. To find a good mapping from image to desired output, we design a parametric model — here based on CNNs — with task-specific components. A loss function is selected that encourages this output, also either explicitly (e.g. Chapter 5) or implicitly together with the appropriate model constraint (Chapter 9). A large annotated training set, annotated either manually (Chapter 7) or with the help of automatic methods (Chapters 8 and 9), guides the search for good model parameters. To verify the outcome of the training process, we evaluate methods on a corresponding test set where the data is assumed to be drawn from the same distribution as the training set. Evaluation metrics summarise performance across the entire test set.

This general approach has lead to massive advances in what automatic visual recognition methods can do. However, one major shortcoming of this approach — which also happens to be the source of its apparent power — is the reliance on statistical correlations in the data. One of the things that makes this problematic is that the learning procedure does not distinguish between relevant and spurious correlations. The latter are artifacts of a particular dataset (Torralba and Efros, 2011), e.g. persons only appear in a specific room (Ionescu *et al.*, 2014) or during certain times of day (Chapter 7). This leads to poor performance when these correlations do not apply — in the so-called out-of-distribution setting. This phenomenon is the subject of much discussion in the recent literature (Arjovsky *et al.*, 2019), and has been referred for example as Clever Hans prediction (Lapuschkin *et al.*, 2019) or shortcut learning (Geirhos *et al.*, 2020).

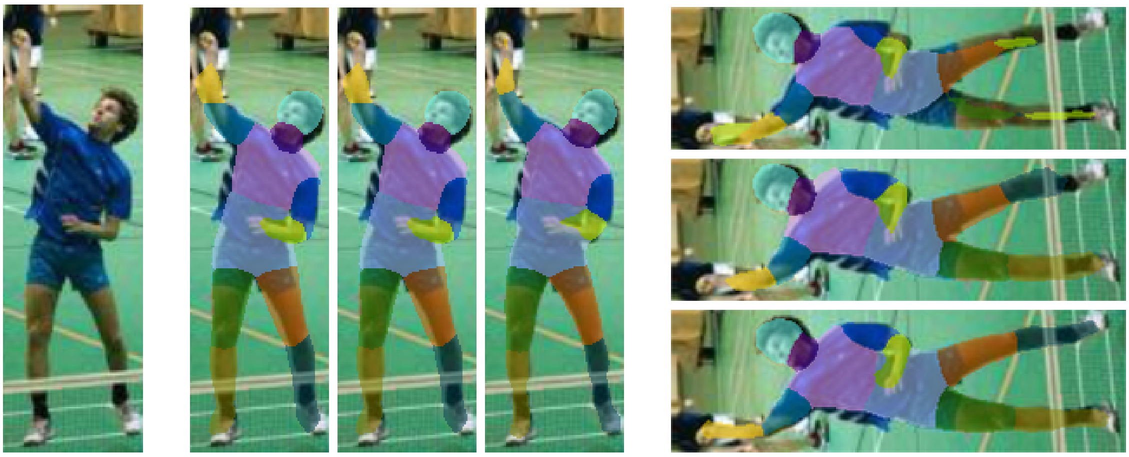
Ultimately, spurious correlations are inescapable in the statistical machine learning setting. No dataset no matter how large will adequately reflect the statistical properties of the real world — not least because the world is constantly changing (Raji *et al.*, 2021). Some aspects of visual perception are also difficult to acquire on the basis of correlations in the data, such as the ability to recognise a familiar object independently of its spatial relation to the viewer. While increasing the size of the training set can address this fundamental inadequacy of datasets as well as of the statistical approach, it will never resolve it. We won't either in this section, but we will outline some promising directions for future work that we believe will help overcome this issue. These are complementary to a fundamentally important reframing of learning in terms of discovering causal rather than purely statistical relations (Schölkopf *et al.*, 2012; **?**).

Before outlining future directions, we will present a couple of failure modes of the method from Chapter 9. These will provide some motivation for the subsequent discussion, which will first focus on benchmarking (Sec. 10.3). Benchmarking plays a central role in modern computer vision — besides also relating to a central contribution of this thesis (Chapter 7). We will argue that guarding against shortcut learning requires improvements to both aspects of benchmarking: dataset design and evaluation. Our main argument is that dataset design in particular should be explicitly guided by the goal of encouraging combinatorial generalisation.

In Sec. 10.4, we will discuss directions for future work on the model side. Here we will focus on the need for better figure-ground organisation as it is relevant to higher-level recognition tasks. Current recognition models are largely static and feedfoward which makes them ill-equipped to respond flexibly to unfamiliar visual inputs. Dynamic models that rely on recurrence and feedback should play a larger role, which will benefit low-level grouping processes. Finally, we will briefly take up some open problems related to detecting and representing people.

10.2.1 Motivating Examples

In Chapter 9 we present a method for 3D human shape and pose estimation that relies on body part segmentation as an intermediate task. The task is to identify pixels belonging to a person and assign to each a body part label. We assume that a single person in the centre of the image is to be labelled. This can thus be viewed as a simplified form of instance segmentation. We use large amounts of data ($> 100K$ examples from a mixture of datasets) and the standard per-pixel cross entropy loss to train models for this task based on a state-of-the-art labelling approach (Chen *et al.*, 2018a). We will demonstrate two types of undesirable behaviour exhibited by our models: (i) failing to generalise to unfamiliar poses not present in the training data, (ii) memorising label noise for specific training examples.



(a) input (b) outputs (c) outputs for rotated inputs

Figure 10.1: We train three different semantic segmentation models with varying degrees of rotation augmentation (up to 30, 60, and 90 degrees respectively). We apply these models to the input image in Fig. 10.1a resulting in the three outputs in Fig. 10.1b. All three models output reasonable labellings with some small amount of degradation for the model trained with the heaviest amount of augmentation. When we rotate the input image by 90 degrees and apply the three models, the difference between the three models becomes clearer.

First, we train three models with varying degrees of rotation augmentation (up to 30, 60, and 90 degrees respectively). In Fig. 10.1, we show some example outputs and observe that all models perform similarly well when a person is upright. Model performance can, however, degrade when the person is viewed from an unusual angle. The model trained without heavy rotation augmentation fails to adequately segment the person spread horizontally across the image. Relatedly, as people are typically upright, common benchmarks will barely distinguish between models that perform differently on such rare instances. In fact, the three models that result in the predictions visualised in Fig. 10.1 preform very similarly on two validation sets (both from the *UP* and *H36M* datasets) in quantitative terms.

Now we consider the second type of undesirable behaviour. The segmentation models we mentioned above are trained on a mixture of datasets including *UP-3D* (Lassner *et al.*, 2017). This dataset is annotated semi-automatically: For each instance, a set of keypoints and the binary silhouette are annotated by hand and an automatic method based on *SMPLify* (Bogo *et al.*, 2016) fits the *SMPL* model to the 2D evidence. Poor fits are discarded but reasonable ones which may contain small errors are kept. Since the semantic labels we use as a target output for our task are derived from these fits they inherit the same errors, such as small misalignments of the limbs that vary from one training example to the next. Remarkably, the model trained to predict these labels has enough capacity to memorise annotation errors specific to individual training examples (see Fig. 10.2) despite the large training set and despite only seeing each example no

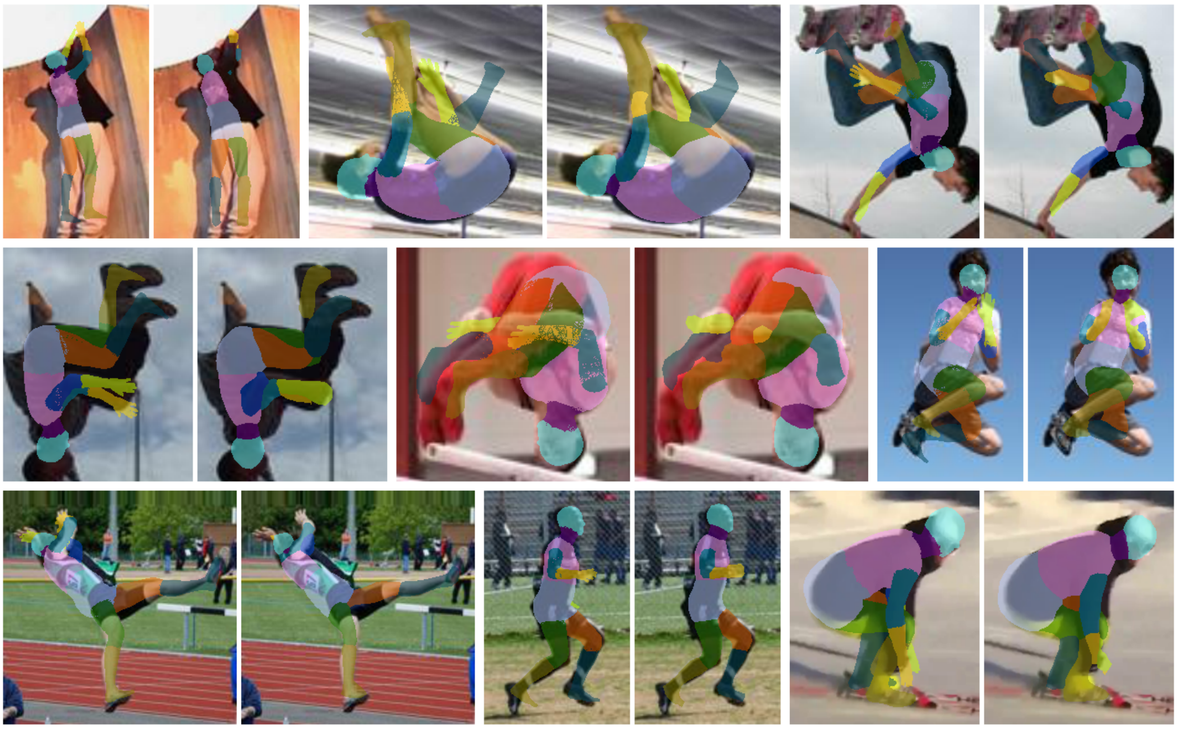


Figure 10.2: A body part segmentation network trained on a mixture of datasets learns to memorise annotation errors for individual training examples. For nine pairs of images, we show the ground truth labelling (left) together with the network prediction (right). We include a mixture of severe annotation errors (first two rows) and more subtle ones such as slightly shifted body parts (last row).

more than 20 – 30 times during training — including with data augmentation (rotation, scaling, and horizontal flipping).

We will refer to these examples in the subsequent discussion.

10.3 Towards Better Benchmarks

Computer vision is a highly benchmark-driven field. For better and for worse, strong performance on standardised benchmarks is critical for the wider adoption of ideas, and drives innovation in models for visual recognition. Accordingly, shortcomings of benchmarks and evaluation practices can have a negative influence on the research priorities of the field. It is thus important to continuously reflect on these shortcomings and address them carefully as a means to guide progress (Paullada *et al.*, 2020). In the following, we will start with a discussion of current dataset collection and evaluation practices before outlining some future directions for work on this topic. The discussion will mainly be informed by the problems we address in this thesis, but to some extent applies more broadly.

10.3.1 Current Benchmarking Practices: An Appraisal

Given the dizzying amount of problems addressed in the field, there is no single recipe for the creation of a benchmark. There are however are some practices and assumptions that broadly apply to most when it comes to data collection and evaluation, as well as common problems that can result from both.

Data Collection

The default assumption underpinning statistical machine learning is that of i.i.d. data. In practice, this means that datasets are created by collecting or recording a large amount of data, and then subsequently splitting this data into training and test sets: either randomly or with loose domain-specific considerations to avoid overlaps between the two sets. This notion of overlap is a very fuzzy one and will mean different things depending on the task and dataset.

In the case of *Cityscapes* (Chapter 7), which was recorded in multiple cities, we split the data by city: both to achieve a balanced distribution by size and geographical location. In terms of separation between training and test set, this means that the exact same scenes and object instances (persons, vehicles, etc.) will most likely not appear in both, and that there will be some variance in architectural styles between the two sets. For *Human3.6M* (Ionescu *et al.*, 2014), a set of actors separately perform a set of pre-defined actions in the same studio. The resulting data is split by subject, and thus requires generalising across person appearance and across different styles of carrying out the same action.

With custom-recorded or -generated data, some mild guarantees can be made for train-test separation. The situation is more complicated when it comes to large web-sourced datasets. Such datasets allow for a larger diversity and scale than is typically feasible with custom data, but with less control when it comes to maintaining a clean split. Metadata, such as the photographer, date, and location, can be helpful in this regard (Lin *et al.*, 2014; Neuhold *et al.*, 2017), but metadata is not always available — especially not in the very large-scale setting involving millions (Russakovsky *et al.*, 2015a) or even hundreds of millions of images (Sun *et al.*, 2017a). In different domains, it has been observed that well-established datasets suffer from overlaps between training and test sets, e.g. by Tatarchenko *et al.* (2019), Barz and Denzler (2020) and Krishna *et al.* (2021) in shape reconstruction, image classification and long-form question answering respectively.

Efforts to automatically detect duplicates (Kolesnikov *et al.*, 2020) can only provide limited guarantees in the large-scale setting. “No duplicates” is also the mildest possible condition for considering two sets of data disjoint, and is not enough to ensure that a task is being solved as intended. Tatarchenko *et al.* (2019) conclude on the basis of a nearest-neighbour oracle that well-performing shape reconstruction methods do not need to perform reconstruction at all. Methods can achieve strong results simply by retrieving similar training set shapes, and they find that many in fact do. Similarly in image classification, (Feldman and Zhang, 2020) show that predictions for several test examples can be traced back to very similar — if not identical — training examples. These results also suggest a partial reliance on a naive retrieval mechanism, which they conclude is necessary for less representative examples that belong to the long tail of the data distribution — and not just necessary for mislabelled examples (Zhang *et al.*, 2017a). These results can be read in several ways, but our takeaway is that a less than carefully designed dataset can obscure, or even encourage, the reliance on shortcuts such as memorisation Tatarchenko *et al.* (2019) and that merely avoiding duplicates — to the extent that it is possible — is not a sufficiently rigorous guarantee; similarity between data points is after all a continuum.

Evaluation Schemes

For many vision benchmarks, performance is reduced to an overall summary statistic and/or curve appropriate for the task, e.g. top-1 or top-5 error, precision-recall curves and mAP, mIoU, PCKh, and MPJPE. Individual test points contribute more or less equally to such aggregate measures. While these measures make it very convenient to compare algorithms, they will naturally reflect performance on the most typical examples in the test set. Without further interventions, they provide little incentive to address corner cases. Many benchmarks thus resort to either implicitly reweighting test examples, or reporting performance for subsets of the data together with the top-line aggregate result.

In certain problem areas, measures are commonly balanced in terms of some discrete semantic attribute, e.g. object class in image classification and object detection (Russakovsky *et al.* 2015a, Chapter 7, Lin *et al.* 2014). This results in additional challenges when the training data has a purposefully lopsided class distribution (Horn *et al.*, 2018; Gupta *et al.*, 2019).

However, long-tailed distributions do not merely occur at the level of class. For example, within the person class some poses are much rarer than others, but benchmarks which treat persons as one class among many, such as the aforementioned, don't consider this aspect. Even benchmarks centred around people often treat all poses — over- or under-represented — equally for the purposes of evaluation, e.g. for pedestrian detection (Zhang *et al.*, 2017b), 2D pose (Lin *et al.*, 2014) or 3D pose (Ionescu *et al.*, 2014; von Marcard *et al.*, 2018). Other characteristics of the data which are relevant for robust recognition and also aren't distributed uniformly include among others: object location and size, occlusion level and occluder type, imaging conditions and capture angle. For most standard benchmarks, an aggregate performance measure that is simultaneously balanced according to multiple such attributes is simply not feasible. This would require overly detailed annotations as well as sufficiently diverse data that covers enough of the relevant combinations.

Given that a scalar performance summary will inevitably end up emphasizing some performance aspect over others, some benchmarks also resort to separately summarising performance for different subsets of the test set. Andriluka *et al.* (2014) for example provide tools for a more detailed analysis of 2D pose estimation methods, reporting performance for different pose clusters, activity types, occlusion levels and truncation levels. In the case of pedestrian detection, several benchmarks provide multiple evaluation settings of varying difficulty based on size and occlusion level (Dollár *et al.*, 2009b; Geiger *et al.*, 2012). For general object detection, Hoiem *et al.* (2012) proposed an early approach to in-depth evaluation of object detectors. These practices have been adopted for the *MSCOCO* object detection benchmark (COCO Analysis Toolkit), and an updated evaluation protocol building on this work was proposed recently by Bolya *et al.* (2020).

While such approaches provide a more nuanced picture of a method's strengths and weaknesses, they are also limited by what we can extract from the available annotations. A detailed quantitative analysis of classification performance on *ImageNet* for example is far from straightforward provided just the image-level labels. This is easier for tasks such as pose estimation thanks to the more fine-grained annotations. But even then, we are limited to analysing performance as it varies in pose space, while having to neglect the other underlying data characteristics mentioned above. Additional annotations can help but are not always practical to acquire. Hoiem *et al.* (2012) for example introduce additional super-class labels for *PASCAL VOC* to be able distinguish between different semantic errors, but Bolya *et al.* (2020) sacrifice this distinction for the sake of applicability to more datasets with more complicated semantic compositions.

An emphasis on quantitative evaluation over a painstaking error-driven and qualitative approach can also easily obscure the reliance on shortcuts. By analysing decision attribution maps, (Lapuschkin *et al.*, 2016) find that a (non-deep) classifier was relying on a source tag present in one fifth of the horse images in the widely used *PASCAL VOC2007* dataset (Everingham *et al.*, 2015).

10.3.2 What Then?

Judging methods by a handful of numbers on large standardised benchmarks is a practice deeply embedded in the culture of the field. In their seminal paper on dataset bias, Torralba and Efros (2011) refer to common laments that "the field is now getting too obsessed with evaluation, spending more time staring at precision-recall curves than at pixels". As this hasn't changed in the intervening decade and is unlikely to change any time soon, it is worth considering ways to make such aggregate performance measures more informative, and we will focus on the question of optimising test set quality without relying on detailed annotations.

Aggregate results on large static test sets — even when these are carefully curated — will nevertheless provide little insight into specific failure modes and can mask a method's brittleness to minor changes in the input. We will argue that work on visual recognition will benefit from dynamic, adversarial evaluation as is common in the robustness literature. By that we mean exploring parameter space in the vicinity of individual test points, as well as seeking out the performance limits of particular models. Analytical tools from the study of human vision, specifically the field of visual psychophysics, can play a role here as well.

Any discussion of evaluation is incomplete without considering the problems that can result from the data collection process. As discussed above, impressive performance on some benchmarks can be achieved while relying on statistical particularities of the underlying dataset instead of performing the task as intended. Addressing this will also require carefully designed datasets with strong experimental controls typically absent from the large i.i.d. setting. We will argue that encouraging combinatorial generalisation is a key goal that can guide dataset design. A broader emphasis on ability-oriented evaluation rather than narrow task-oriented evaluation also matters. We will discuss these terms in greater detail later in this section.

Both the dynamic evaluation and experimental controls we will argue for are very difficult to achieve with natural data. While overcoming this difficulty poses a research challenge in its own right, we believe that synthetic datasets should play a larger role in the areas we address here, in which natural data has traditionally dominated in benchmarking. This would also help address the difficulty of test set selection without access to detailed annotations.

10.3.3 Improving Aggregate Performance Measures

Given a fixed test set, how do we gain more nuanced quantitative insights about model performance? Above we mentioned evaluation schemes that rely on splitting the dataset according to different domain-relevant factors. However, these require detailed annotations that are typically not sufficiently exhaustive if at all available. Are there then alternatives for selecting informative subsets of the data for evaluation?

One simple approach that is surprisingly uncommon would be to report worst-case performance using pre-determined percentages of the test data. This would provide some information on a method's reliability that might otherwise be obscured by an average score. Large i.i.d. test sets often suffer from redundancy, often containing very similar test points that water down the results and bias the comparison between methods. A method that performs well across a wide range of conditions will be penalised relative to a method that excels on average but performs poorly on difficult test cases.

Along these lines, recent work has sought to identify fixed "difficult" subsets. Li and Vasconcelos (2019) propose to remove examples that are biased towards an unwanted representation. They apply this to action recognition data by identifying examples that are easily classified with single-frame cues. However, this type of "representation bias" is difficult to define in a problem-agnostic manner and can involve overly strong assumptions. Bras *et al.* (2020) instead suggest a general criterion, namely how often a test point is correctly identified under different train-test splits. They propose a simple heuristic algorithm to efficiently remove examples with a high "predictability" score.

Influence functions represent a technique from robust statistics that has been adopted for model interpretability (Koh and Liang, 2017), and which might be useful for the purpose of evaluation. In the interpretability setting, these capture the influence of training points on model predictions. They could be used to derive difficulty scores for individual test points, especially since we know that memorisation plays a large role in contemporary deep learning (Zhang *et al.*, 2017a; Arpit *et al.*, 2017; Feldman, 2020; Feldman and Zhang, 2020). In some problem domains, detailed annotations could also be used to derive difficulty scores, e.g. in 2D/3D pose estimation and 3D reconstruction. Test examples could be assigned weights based on their distance to the training set, possibly also taking training-time augmentations into account.

An alternative to seeking difficult subsets is identifying a balanced test set that contains a mixture of easy and hard examples. This could be used to improve existing test sets rather than merely complementing them. There is limited work on this problem but it deserves more attention. Balduzzi *et al.* (2018) take a step in this direction with the goal of ranking a set of agents that each perform a set of tasks or compete against every other agent. Their proposed evaluation scheme relies on finding the latent structure in performance tables (agent-vs-task or agent-vs-agent). With this they addresses a number of shortcomings they identify with performance averages or ELO rankings, including sensitivity to redundant tasks — discussed above. Contrary to what standard evaluation schemes suggest, they find that reinforcement learning agents had not in fact yet outperformed humans on widely-used Atari benchmarks.

Psychometrics, the field of study devoted to measuring latent abilities and traits of humans, can provide useful tools for evaluating algorithms. One particular methodology in this area, item-response theory, has found limited use in machine learning (Lalor *et al.*, 2016; Martínez-Plumed *et al.*, 2019). This involves fitting statistical models to the "responses" of different test participants — predictions of recognition models — to a set of "items" — test examples. This methodology allow us to estimate item characteristics such as difficulty and suitability for discriminating between participants with the goal of arriving at more informative test sets and rankings. Possible extensions to this approach can perhaps incorporate information on the training set to which we have full access unlike with human subjects.

10.3.4 Dynamic, Adversarial Evaluation and Synthetic Benchmarks

To better understand the shortcomings of recognition models, it is imperative to move beyond evaluating them on static test benchmarks. In the robustness literature, this is common practice (Carlini *et al.*, 2019). To probe model robustness, one seeks the smallest possible additive change to the image that induces a false prediction by maximising the loss w.r.t. the image. When no restriction is placed on such a corruption other than a bound on its norm, these are then referred to as adversarial perturbations and are often imperceptible to humans. While pixel space perturbations still pose a very challenging problem, there is more room for work on robustness in world space or object space.

Simple geometric transformations (e.g. rotations and translations of the full image) can confound models as shown by Engstrom *et al.* (2019), who similarly use an adversarial approach to discover small changes that cause image classification methods to fail. Liu *et al.* (2019b) propose the concept of a parametric norm-ball. Rather than considering distance in pixel space they use a differentiable renderer to consider robustness in the space of parameters, such as lighting and shape distance. Similarly, Alcorn *et al.* (2019) use a differentiable renderer to adversarially discover out-of-distribution poses for rigid objects. Shetty *et al.* (2020) train a model to perform a type of adversarial image editing. This modifies the texture of objects while keeping shape and scene composition intact.

Synthetic datasets — given that they grant complete control over the underlying scene parameters — can enable the kind of rigorous and flexible evaluation we have discussed so far. Advances in generating synthetic datasets (Richter *et al.*, 2016; Qiu and Yuille, 2016; Wrenninge and Unger, 2018; Roberts and Paczan, 2020; Devaranjan *et al.*, 2020; **?**) as well as in differentiable rendering (Nimier-David *et al.*, 2020; Laine *et al.*, 2020; Kato *et al.*, 2020) allow for generating more realistic adversarial benchmarks, for people detection and analysis especially. Besides providing easy access to the underlying scene variables, they also have the benefit of making it easier to generate annotations that are tedious if not very difficult to generate manually, such as pixel masks and continuous 3D quantities.

Here it is important to note the following: There is a distinction between generating realistic surface appearance (Wrenninge and Unger, 2018; Roberts and Paczan, 2020) and generating realistic scene compositions (Devaranjan *et al.*, 2020; Hassan *et al.*, 2021) and both represent separate, if related, challenges, whose relative importance depends on the goal. For evaluating recognition methods, realistic and varied scene compositions as well as potentially differentiable control over scene variables can result in challenging benchmarks even without realistic surface appearance (Zitnick *et al.*, 2016).

Control over scene variables can also allow for a more robust analysis of model behaviours, similar to how human vision is studied using the tools of visual psychophysics (Lu and Dosher, 2013). Methods and insights from this field can be adopted to rigorously characterise failure modes beyond merely finding specific points of failure or generating individual difficult test examples. Recent efforts along these lines include the work of Wichmann *et al.* (2017); RichardWebster *et al.* (2018a,b).

10.3.5 Towards Combinatorial Robustness and Ability-Oriented Evaluation

So far we have mainly focused on evaluation schemes, but designing the right training and test sets is of equal importance. Naturally, the details will depend on the nature of the problem being considered but we will attempt to articulate a principle here that should apply more generally. Our view is that dataset design should be explicitly guided by the goal of encouraging combinatorial robustness. By this we mean the ability to generalise to unseen combinations of known parameter settings. This is a broad definition that encompasses a wide range of challenges relevant to recognition. Before enumerating a number of concrete examples, let us first attempt to draw a distinction between combinatorial generalisation and domain adaptation.

Defining Combinatorial Robustness

Let there be two sets of images: X_{train} and X_{test}. We assume that each image $\mathbf{x}$ is the result of a complicated generative process which operates on a set of latent variables $Z = \{\mathbf{z_1}, ..., \mathbf{z_N}\}$ (alternatively: "factors of variation" Bengio 2009). These can include continuous and discrete quantities such as sensor properties, lighting conditions, the weather, what objects are present, properties of these objects such as their appearance or pose, their spatial configuration in the scene, and numerous others. They need not be independent and for any given recognition task we will be only interested in recovering a strict subset, e.g. the class of the dominant object in the image. We will furthermore assume that for any arbitrary subset of the latent variables $Z_s \subseteq Z$, the values of its members will be jointly distributed according to $p_{train}(Z_s)$ and $p_{test}(Z_s)$ in X_{train} and X_{test} respectively.

In the traditional domain adaptation setting, the assumption is that the individual *marginal* distributions of one or more parameters will undergo a shift from training to test set. Often in fact, a pair of marginal distributions $p_{train}(z_i)$ and $p_{test}(z_i)$ will have non-overlapping mass. This means that the test set is designed such that all its elements share some underlying characteristic(s) not present in the training set, e.g. captured with a different sensor (Saenko *et al.*, 2010), synthetic rather than real (or vice versa) (Peng *et al.*, 2018), corrupted in a specific manner (Hendrycks and Dietterich, 2019), captured under different weather conditions (Chen *et al.*, 2018c), contains different sub-classes (Santurkar *et al.*, 2021), and others.

In the case of combinatorial generalisation, there need not be such a restriction on the marginal distributions of individual factors; they may even be equivalent across train and test sets. Instead, what matters is that the *joint* distribution of a subset of the factors undergoes a controlled shift between training and evaluation, e.g. $p_{train}(z_i, z_j)$ vs. $p_{test}(z_i, z_j)$. This means that the test set should contain unseen combinations of possibly familiar latent factor values. We will refer to the ability to handle such shifts as *combinatorial robustness w.r.t. some property of the data.* Depending on the setting and on what latent parameters are assumed, this can result in a very diverse set of challenges, and we will now discuss examples to clarify these definitions.

Examples

One well-studied problem that requires combinatorial robustness is the recognition of objects in unfamiliar contexts. If we assume a simplified image classification setting where we have a limited set of object instances and scenes, the relevant factors would be z_{class} and z_{bg} respectively. That current models struggle with unfamiliar combinations of these factors has been often demonstrated, e.g. in Rosenfeld *et al.* (2018) and Shetty *et al.* (2019). Some datasets have been proposed recently to study this problem from a robustness perspective. Sagawa *et al.* (2020) present a synthetic dataset that mixes and matches two species of birds (waterbirds and land birds) against two types of backgrounds (land and water). The combinations are distributed differently in training and test sets. Xiao *et al.* (2021) use automatic segmentation methods to generate challenge sets from *ImageNet* in a similar fashion.

Other examples that are highly relevant to recognition but are not commonly studied in isolation would be combinatorial robustness w.r.t. object scale, rotation, or scene illumination. Consider a training set for object classification in which objects occur at a wide range of scales, but where specific classes are restricted to separate scale ranges for training and evaluation. Evidence suggests that current models would struggle to handle such a shift (Engstrom *et al.*, 2019; Alcorn *et al.*, 2019). Robustness to geometric transformations such as scale and rotation are currently addressed in a data-driven manner via heavy data augmentation (see Fig. 10.1). This robustness is naively encoded in the model in the form of redundant weights for different cases encountered during training. In top performing models, there is no mechanism for handling unseen sizes and orientations in an object-agnostic manner, and standard benchmarks provide little incentive to develop such mechanisms. Benchmarks like *Caltech* (Dollár *et al.*, 2012b) and *MSCOCO* (Lin *et al.*, 2014) do report performance for different object scale ranges, but without explicit controls for what is seen during training. To be clear, scale handling is challenging even without the pursuit of combinatorial robustness, as we for example discuss in Chapter 2. However, datasets with such controls are necessary to drive further modelling innovation and avoid the heavy reliance on data that is customary at present.

Similarly, we can imagine an object detector trained on images captured under a wide range of illumination conditions, but with certain objects only showing up in poorly-lit scenes during evaluation. There is a lot of evidence for texture bias in state-of-the-art recognition methods (Baker *et al.*, 2018; Geirhos *et al.*, 2019). Objects appear to us as the result of interactions between light, viewpoint and material-dependent properties. If current models do not implicitly learn to perform a kind of intrinsic image decomposition (Barrow and Tenenbaum, 1978) — which they most likely don't, they will end up learning biased statistics of object appearance and be unable to generalise accordingly. It should be noted that scene illumination also affects the ability to perceive edges and surfaces thus affecting models that may rely more on shape cues than texture (Tuli *et al.*, 2021).

Combinatorial robustness is naturally relevant to the detection of multiple objects. It also matters at the level of individual objects as it pertains to e.g. (i) recognising objects consisting of unfamiliar configurations of known parts, (ii) occlusion handling with diverse occluders and levels of occlusion, and (iii) recognising familiar poses for objects with unfamiliar appearances. This especially applies to highly articulated objects e.g. the human body, but targeted evaluations of the kind we describe here are rare.

One example of this can be found in Lehrmann *et al.* (2013), where they propose a structured prior over human pose. In one experiment, they learn the prior on a set of standing poses where at most one arm is raised in any given example. They show that in contrast to competing models, theirs can produce samples where both arms are raised. The model — despite not seeing such examples during training — has learned from the data that the spatial correlation between arms is weak and generalises accordingly. It's not obvious that modern pose priors based on generic deep generative models (Pavlakos *et al.*, 2019a; Xu *et al.*, 2020), which model pose in a global manner, can exhibit this kind of generalisation. Similarly, contemporary methods for pose estimation will possibly struggle if an artificial correlation between pose and appearance is introduced into the training set, e.g. if trained on a mixture of real humans and humanoid-like figures, where each subset covers a different part of pose space.

Ability-oriented Evaluation

As we pointed out, many of these problems have been studied in some form or the other. Especially in recent work however, they are often addressed from a separate robustness perspective: The goal is either to shed light on a problem (Engstrom *et al.*, 2019; Alcorn *et al.*, 2019; Xiao *et al.*, 2021) or to propose a specialised solution (Sagawa *et al.*, 2020) that is rarely adopted outside of the robustness literature.

This disconnect between work on model robustness and work on specific recognition tasks can be partly bridged with newer benchmarks that don't focus on narrow task performance in the large i.i.d. setting. Combinatorial robustness to changes in background, scale and illumination for example are arguably basic requirements for robust recognition, but are not measured explicitly by standard benchmarks. What we are in sense then also arguing for here is what (Hernández-Orallo, 2017) refers to as "ability-oriented" vs. "task-oriented" evaluation (there albeit in reference to higher-level cognitive "abilities" compared to the more basic ones we discuss here). We should also emphasize that we think both approaches are complementary.

While not explicitly focused on combinatorial robustness, a similar approach is starting to gain traction in natural language processing (NLP): one that focuses on specific linguistic competencies that are common to different tasks. One recent example is Checklist (Ribeiro *et al.*, 2020), in which model evaluation is approached from a software-testing perspective. A battery of tests that cover different behaviours or abilities (e.g. synonym handling, robustness to typos) is designed to evaluate models in a more targeted manner, as a complement to task performance. This would be a useful approach to problems like pose estimation and object detection but with carefully designed controls on the training side.

Related Work

To conclude this section, we would like to discuss some connections to recent work.

The importance of pursuing the goal of combinatorial generalisation has been articulated elsewhere, e.g. by Battaglia *et al.* (2018). There, however, this goal is cited to motivate the adoption of graph neural networks. These models exhibit some forms of combinatorial robustness by virtue of the inputs they require and how these are processed: partly as separate nodes and edges. As they acknowledge, this does not address the challenging problem of extracting such graphs from raw inputs to begin with, e.g. images.

A related problem is the unsupervised learning of disentangled representations (Schmidhuber, 1992; Desjardins *et al.*, 2012; Higgins *et al.*, 2018). There the goal is to learn a generative model of images such that a set of relevant latent factors are discovered automatically and made separable in each image representation. Such methods are typically applied to simple synthetic datasets with a limited number of factors. While no assumptions are made on the type of factors in advance, recent approaches rely on an unrealistic inductive bias in the form of a strong independence assumption (Higgins *et al.*, 2017; Kim and Mnih, 2018). The datasets used for training contain every possible combination of factors and thus no nuisance correlations exist. (Exceptions that relax this assumption include the work of Bozkurt *et al.* (2019) and Träuble *et al.* (2020), who find that standard approaches struggle as a result.)

It is our contention, however, that stronger inductive biases are required on the model side. These will also most likely be very different for different factors of variation, and not be discoverable with a generic CNN-based architecture and the right loss function. Instead, carefully designed datasets focusing on different "abilities" are required to spur the development of the corresponding architectural features. This is a more specific version of the argument made by (Locatello *et al.*, 2019) who, based on both theoretical and empirical results, show that more inductive biases are necessary in this setting — even with the strong independence assumption. This argument is also aligned with work on unsupervised object-centric representation learning (van Steenkiste *et al.*, 2018; Burgess *et al.*, 2019; Greff *et al.*, 2019; Locatello *et al.*, 2020), where an architectural bias towards handling multiple objects is shown to be useful for the generative modelling of corresponding data.

A highly related concept is the assumption of "independent causal mechanisms" from the causal machine learning literature (Schölkopf *et al.*, 2012). To discover the causal structure that underlies the data generation process — as opposed to merely learning surface correlations — it is assumed that the data is generated by a set of independent "mechanisms". The learning approach is set up such that these can be discovered. Parascandolo *et al.* (2018) apply this principle to a modified version of *MNIST* with an appropriate architectural inductive bias and learning objective: A set of CNNs are adversarially trained to each specialise in undoing a simple transformation, such as some spatial shift, colour inversion, or denoising. We argue that handling more challenging transformations on more natural data will require more challenging benchmarks as well as more complicated architectures, especially when multiple such transformations are composed arbitrarily.

The problem of generalising to unseen combinations of known elements is the subject of much work in NLP, as well as at the intersection of vision and language (Johnson *et al.*, 2017). This is typically referred to as compositional (Keysers *et al.*, 2020) or systematic (Lake and Baroni, 2018; Bahdanau *et al.*, 2019) generalisation based on related linguistic concepts. A complete treatment of the extensive related work is beyond the scope of our discussion, but a couple of contributions are particularly relevant. Hupkes *et al.* (2020) define different types of "compositional behaviour" to address inconsistencies in the literature. Some of their definitions are directly applicable to vision problems. Keysers *et al.* (2020) propose a metric and algorithm for automatically generating datasets from natural language data requiring varying degrees of combinatorial robustness.

Applying such a method to visual data is not straightforward. In the case of language it relies on measuring word distributions which in principle can be obtained reliably via tokenisation. Obtaining different visual elements of interest on the other hand from natural images is extremely difficult. In general, creating controlled datasets is a key challenge, especially with natural data. Thus we think synthetic data should in general play a larger role as argued above. However, there are also efforts to collect such data in the real-world. Barbu *et al.* (2019) propose a smart data collection protocol that allows them to gather data from real-world indoor scenes while controlling for various variables such as object pose and background. While this is intended for use as an image classification test set, such an approach could potentially be used to create challenging train-test set pairs.

Finally, we started this discussion by drawing a contrast between the problem of combinatorial robustness and domain adaptation in a formal sense. In an informal sense, domain adaptation requires of the model a "blind leap" between training and test data. A model for example that has only learned to recognise objects on the basis of natural texture is expected to generalise to synthetic texture without having been exposed to the latter. This is particularly challenging given the strongly data-driven nature of current standard models. Instead, with a dataset designed to encourage combinatorial robustness the requirement is less stringent, and can thus perhaps direct model development in a more targeted fashion.

10.4 Towards Better Models

In the previous section we discussed issues surrounding benchmarking. These were primarily motivated by the larger goal of designing stronger recognition models. In Sec. 10.3.5, we outlined some specific problems and desiderata that can inform dataset design in service of this goal.

In this section, we will discuss some specific future directions regarding model design itself. In the first part, we will focus on the need for dynamic models that use recurrence and feedback. We will motivate this primarily via low-level vision problems. These, as we've argued above, are relevant to every recognition task we address here.

10.4.1 Dynamic Models with Recurrence and Feedback

Models that currently dominate visual recognition — based on CNNs — are largely static and feedforward. By static we mean that models consist of filters whose weights and support remain fixed after training[10], i.e. are applied as is to every subsequent input without adaptation. This descriptor also applies to recurrent units, e.g. LSTMs, when these are optimised and applied for a fixed number of iterations. By feedforward, we mean that visual processing occurs in a hierarchical manner, successively extracting low-level to high-level features. This describes most successful models for image classification (e.g. Simonyan and Zisserman 2015, He *et al.* 2016). For tasks with a localisation component, some commonly used models depart from this in a limited fashion (Lin *et al.*, 2017b; Tan *et al.*, 2020). In these models features are combined in a top-down manner, still however, with fixed operations. There are models that depart more dramatically from the static feedforward paradigm. However, these have not yet attained either widespread adoption or meaningfully better performance as measured on standard benchmarks.

Nonetheless, we think that models that incorporate dynamic recurrence and feedback should play a larger role in recognition as we will subsequently argue. For a wider-ranging discussion on potential computational benefits of recurrence in artificial vision models, we recommend the position paper of van Bergen and Kriegeskorte (2020). We will present some additional motivation here, and discuss some recent work that points to fruitful research directions.

In Sec. 10.3.5 we argued that standard models are not sufficiently flexible when it comes to certain basic abilities relevant to recognition, e.g. being able to recognise a familiar object when it appears in unfamiliar conditions such as illumination or distance to camera. When it comes to recognising some object at different scales, using static operations has the following consequences: (i) It requires redundant model weights for different scales, (ii) information on scale is inseparable from information on object appearance, (iii) and the model does not generalise beyond scales encountered during training. Scaling, like some other transformations (e.g. rotation, Fig. 10.1) or corruptions (e.g. noise and blur) can be approximated with the repeated application of simple operations. By repeatedly rescaling or rotating an image by a small amount, or by repeatedly applying a small amount of noise, we can approximate an arbitrarily scaled or noisy image (see Fig. 10.3).

Current recognition models evidently have the capacity to learn to “undo” such operations in some form before classifying the underlying object (Rusak *et al.*, 2020). However, these models have no built-in inductive bias to learn simple inverse operations that can be repeatedly invoked in a manner that allows for generalising to unseen degrees of transformation or corruption. Instead, we must rely on having the right training distribution (see Fig. 10.3). Furthermore, with static networks the same computational budget is allocated to images with very different characteristics and which might require very different amounts of processing. For example: Detecting edges in uncorrupted

[10] This also applies to “weight-less” operators such as max-pooling, which have fixed support.

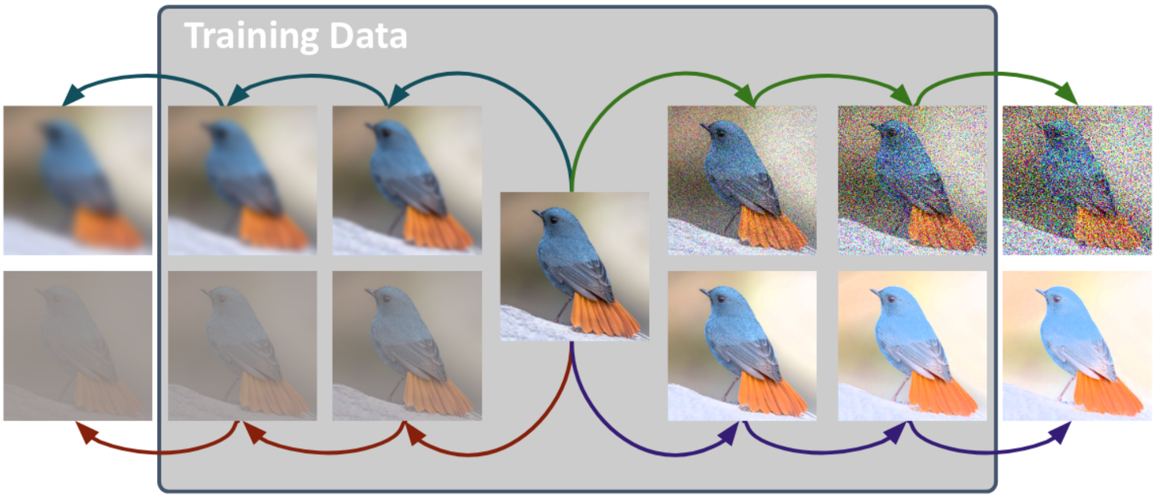


Figure 10.3: Many image transformations that can occur in practice, such as noise, blur, changes in contrast and brightness (depicted above) but also geometric transformations such as scaling and rotation, can be approximated by the repeated application of simple operations. While current recognitions models can learn to handle these given the right data distribution, they struggle when faced with an unseen transformation. Models with inductive biases that allow for learning and recursively applying the inverses of such operations might overcome this limitation.

images is more straightforward than in noisy images. Reliable detection in the latter case is more dependent on being able to integrate neighbourhood statistics or applying the proper smoothing in an structure- or edge-aware manner, which in classical approaches requires longer processing (Weickert, 1998; Mrázek and Navara, 2003; Roth and Black, 2009).

A promising line of work that can help address these problems involves so-called implicit (Bai *et al.*, 2019) — or alternatively declarative (Gould *et al.*, 2019) — layers. These are layers which are specified in terms of an objective rather than a fixed sequence of computations. Examples include differentiable layers that compute a fixed point (Bai *et al.*, 2019), or solve a differential equation (Chen *et al.*, 2018b) or convex minimisation problem (Agrawal *et al.*, 2019). Key to making this work is directly differentiating the solution w.r.t. the layer inputs, as opposed to differentiating through each computational step. This enables layers that can carry out powerful variable-length computations with a constant memory requirement during training.

One recent approach related to this idea is the method of Linsley *et al.* (2020). They apply a recurrent neural network (RNN) to a contour following task, where the goal is to segment out a simple contour among several distractors. The conventional approach to training RNNs, “backpropagation through time“ (BPTT), requires unrolling recurrent computations to a fixed number of steps, which comes at a high memory cost. More importantly, Linsley *et al.* (2020) show that the RNN performs optimally at the same number of unrolling steps but degrades afterwards. As a result, a conventionally-trained RNN does not generalise to longer contours than those seen during training. When using implicit gradients on the other hand, the RNN generalises ably to longer inputs in contrast to RNNs optimised for a fixed number of steps and in contrast to CNNs, which have fixed depth. They additionally integrate recurrent units into a *Mask R-CNN* network (He *et al.*, 2017), and show that this segments objects with a gradual flood-filling approach while matching the performance of the baseline *Mask R-CNN* on the *MSCOCO* instance segmentation task. While this is not explicitly argued, such an approach has the potential to enable generalisation to unseen object configurations in the same way it enables length generalisation on the contour following task.

There are several possibilities for building on this work, and we will briefly discuss a few here: (i) exploring the design space of such models, (ii) focusing on improving the robustness of early CNN layers to low-level corruptions, (iii) integrating control mechanisms for implicit layers possibly recruiting high-level information, and also (iv) using such approaches for fitting complicated shapes to data, e.g. articulated human body models.

Linsley *et al.* (2020) show that the behaviour of known recurrent units (Linsley *et al.*, 2018) can change when trained in a manner that allows for more flexible computations. These units have fixed weights and support, and one could revisit or extend more complicated neural network components, especially ones that allow for geometric transformations of filter weights, e.g. Spatial Transformer Networks (Jaderberg *et al.*, 2015) and Deformable Convolutions (Dai *et al.*, 2017), or adjustments to the weights themselves, as in e.g. Dynamic Filter Networks (Jia *et al.*, 2016). These methods rely on learned adaptation functions that are similarly static and dependent on the training distribution. More flexible adaptation of the filters in a recurrent manner can result in more powerful behaviour.

A further area of focus could be improving early layers of CNNs to be more robust to low-level corruptions. Some recent work on robustness deals with so-called ”common corruptions“ (see e.g. the *ImageNet-C* benchmark from Hendrycks and Dietterich 2019). These are modifications in pixel space that resemble named image corruptions encountered in practice, such as specific types of noise and blur artifacts, weather effects or colour distortions. Many methods addressing these corruptions focus on improvements to the training data (Rusak *et al.*, 2020) and it’s unclear how combinatorially robust such approaches are (see Sec. 10.3.5). Instead, architectures with specially designed low-level processing modules might be able to address this problem more robustly.

One line of work relevant to the above involves reformulating classical approaches to image restoration (Rudin and Osher, 1994; Roth and Black, 2009) as learning problems with modern deep architectures (Chen and Pock, 2017; Kobler *et al.*, 2017; Effland *et al.*, 2020; Kobler *et al.*, 2020). These models still largely rely on a fixed number of layers. One could leverage advances in training implicit layers to obtain models that can process images more flexibly, and also train models that can handle different types of restoration tasks as opposed to training corruption-specific or even corruption level-specific models.

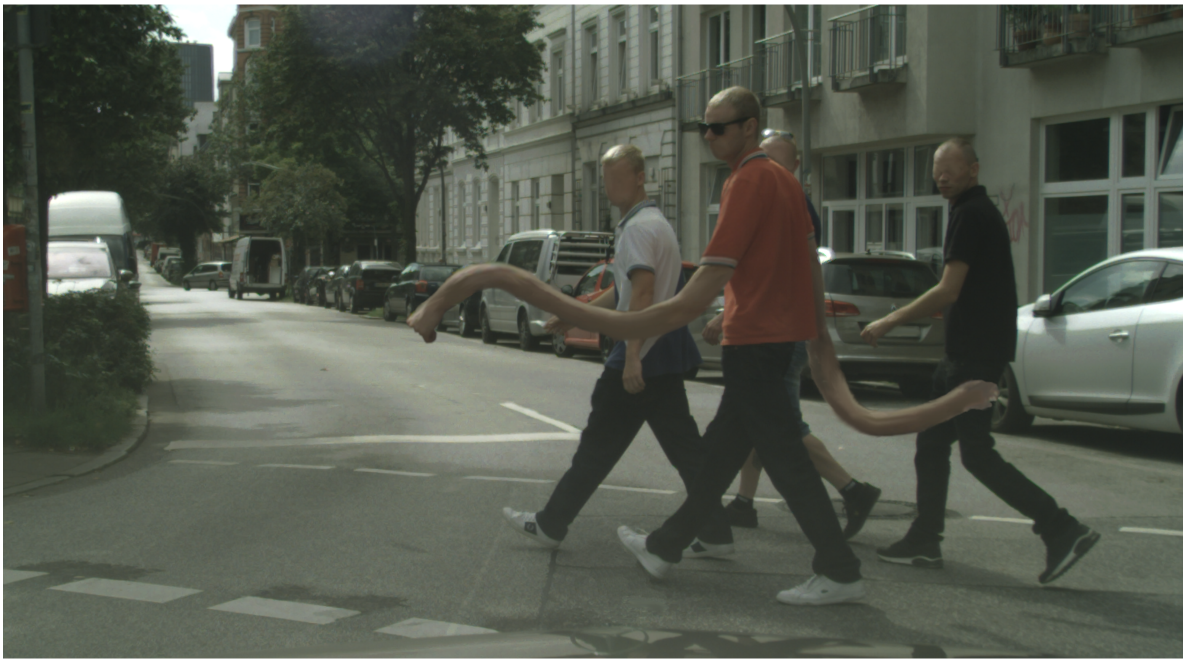


Figure 10.4: How would current 3D shape and pose estimation approaches handle this artificial example? This is a useful way to think about their limitations at different levels, whether prediction or modelling.

There is also some redundancy in these models: See Figs. 11 and 12 in Effland *et al.* (2020), which show the learned filters and activation functions corresponding to one denoising and one deblurring network. The filters for both are remarkably similar but the activation functions aren't. More sophisticated network designs with different activation pathways (Goodfellow *et al.*, 2013) can make use of this fact.

Another exciting possibility is designing mechanisms for controlling the execution of implicit layers with high-level information through feedback connections. Similar ideas have been previously explored with traditional architectures, e.g. by deriving some control signal from the classifier confidence (Spoerer *et al.*, 2020). Controlling low-level processing modules such as the above with high-level information, e.g. learning to suppress nuisance information, will in turn lead to more robust high-level recognition and also to advances in image restoration.

Flexible computation is useful not just for low-level vision, but also for object recognition. Models such as *Capsule Networks* (Sabour *et al.*, 2017) use iterative computations to explain object appearance dynamically rather than through the use of fixed templates. This kind of dynamic inference would be particularly beneficial when estimating the pose of articulated objects such as people, and make methods less dependent on having access to the right distribution of poses during training.

To emphasize this point, we consider an artificial example (Fig. 10.4) and its implications for current 3D human body recovery methods. Humans have no problem understanding the admittedly unlikely pose and shape, but for automatic methods it would cause problems at multiple levels: ”Bottom-up“ predictors that extract 2D keypoints or part segmentations would struggle. Even though these produce pixel-wise predictions, evidence suggests that their predictions are strongly correlated (Fig. 10.2) and dependent on the pose distribution (Fig. 9.4) at training time. This will naturally cause problems for lifting to 3D whether through prediction or fitting. Methods that attempt to predict the parameters of a body model would naturally struggle, not least because current parametric models (e.g. *SMPL*) could not accurately represent this pose. Addressing this will at the very least require pixel-wise methods that are more sensitive to low-level evidence and more tightly integrated with flexible object-level reasoning that can adjust object models to fit unfamiliar inputs. Developing more advanced human body models to support this flexible inference also poses a significant challenge.

10.4.2 Towards Stronger Object Detection

We conclude with a short discussion of our results on pedestrian detection. What generally stands out both from these results as well as from the subsequent literature on detection is that certain key things have changed very little despite the wealth of work on the subject: The fundamental approach to detection has remained the same. When it comes to performance, the primacy of the training data as well as having strong but conceptually simple image representations also matters significantly.

Already in Munder and Gavrila (2006) it was observed that “[t]he greatest performance gain was, however, achieved by increasing the training sample size” when comparing different feature and classifier types on the task of pedestrian classification. The benefits of using larger training sets have been observed repeatedly since in detection, e.g. in Nam *et al.* (2014) with hand-crafted features and decision forests, and similarly in Chapter 4 with our cross-dataset generalisation experiments. This is all the more true for CNN-based methods that are able to benefit from even more large amounts of data. We demonstrated this for example in Chapter 5, as have subsequently Zhang *et al.* (2017b) and Braun *et al.* (2019). They both show that pre-training CNN-based detectors on their respective large-scale datasets gives a performance boost on smaller ones.

Recent state-of-the-art detectors, e.g. (Tan *et al.*, 2020), still demonstrate the success of these basic ingredients together with the sliding-window approach to detection. We depict this approach in Chapter 2 (Fig. 2.3) as a multi-dimensional grid-labelling problem, where each grid point typically represents a spatial location and scale and, by implication, a sub-area of the image. The labelling is such that sub-areas of the image either contain or do not contain an object of interest. Detectors are trained to separate these groups of sub-images, and additionally to refine the location of object hypotheses.

Despite the successes of this approach as evidenced by strong benchmark performance, there are some problems with it. Some of them stem from modern DNNs’ ability to effectively minimise the learning objective via memorisation if need be (Zhang *et al.*, 2017a). The objective imposes an artificial, binary separation of feature vectors into objects and non-objects, and current networks are capable of confidently learning these labels even in more ambiguous cases. Some evidence of this in the context of object detection can be seen in Jiang *et al.* (2018) (Fig. 2a). This shows that the network to a significant degree successfully learns how to separate positives from negatives with perfect confidence for many training examples regardless of how much (or how little) they overlap with ground truth boxes. There is a lot of work that attempts to propose improvements to the objective function, e.g. the aforementioned paper. We believe that here, as above, stronger low-level processing with dynamic models will help produce image representations that can better separate foreground from background. These will benefit from explicit exchange of information not merely at the window level (Rothe *et al.*, 2014) but also at earlier stages of the processing pipeline, in contrast to the late integration of redundant decisions that is common in current methods.

Bibliography

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng (2016). TensorFlow: A System for Large-Scale Machine Learning, in *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*. Cited on page 167.

A. Agarwal and B. Triggs (2004). 3D Human Pose from Silhouettes by Relevance Vector Regression, in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*. Cited on pages 30 and 34.

A. Agrawal, B. Amos, S. T. Barratt, S. P. Boyd, S. Diamond, and J. Z. Kolter (2019). Differentiable Convex Optimization Layers, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Cited on page 195.

P. Agrawal, R. B. Girshick, and J. Malik (2014). Analyzing the Performance of Multilayer Neural Networks for Object Recognition, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*. Cited on pages 76 and 86.

I. Akhter and M. J. Black (2015). Pose-conditioned joint angle limits for 3D human pose reconstruction, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 31, 45, 46, 47, and 173.

M. A. Alcorn, Q. Li, Z. Gong, C. Wang, L. Mai, W. Ku, and A. Nguyen (2019). Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 188, 190, and 191.

B. Alexe, T. Deselaers, and V. Ferrari (2012). Measuring the Objectness of Image Windows, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34(11), pp. 2189–2202. Cited on page 15.

M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele (2014). 2D Human Pose Estimation: New Benchmark and State of the Art Analysis, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 30, 32, 169, and 185.

M. Andriluka, S. Roth, and B. Schiele (2009). Pictorial structures revisited: People detection and articulated pose estimation, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. Cited on page 39.

M. Andriluka, S. Roth, and B. Schiele (2010). Monocular 3D pose estimation and tracking by detection, in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. Cited on page 44.

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis (2005). SCAPE: shape completion and animation of people, *ACM Trans. Graph.*, vol. 24(3), pp. 408–416. Cited on pages 29 and 34.

P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik (2011). Contour Detection and Hierarchical Image Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33(5), pp. 898–916. Cited on pages 139, 141, 142, 143, and 147.

P. A. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marqués, and J. Malik (2014). Multiscale Combinatorial Grouping, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 135.

S. Ardeshir, K. M. Collins-Sibley, and M. Shah (2015). Geo-semantic segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 126.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz (2019). Invariant Risk Minimization, *CoRR*, vol. abs/1907.02893. Cited on page 180.

A. Arnab, C. Doersch, and A. Zisserman (2019). Exploiting Temporal Context for 3D Human Pose Estimation in the Wild, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 33.

D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien (2017). A Closer Look at Memorization in Deep Networks, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Cited on page 187.

H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson (2015). From generic to specific deep representations for visual recognition, in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 67, 75, and 86.

V. Badrinarayanan, A. Kendall, and R. Cipolla (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39(12), pp. 2481–2495. Cited on pages 122, 130, 132, and 134.

D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. C. Courville (2019). Systematic Generalization: What Is Required and Can It Be Learned?, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. Cited on page 193.

S. Bai, J. Z. Kolter, and V. Koltun (2019). Deep Equilibrium Models, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Cited on page 195.

N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman (2018). Deep convolutional networks do not classify based on global object shape, *PLoS Comput. Biol.*, vol. 14(12). Cited on page 190.

A. O. Balan, M. J. Black, H. W. Haussecker, and L. Sigal (2007a). Shining a Light on Human Pose: On Shadows, Shading and the Estimation of Pose and Shape, in *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*. Cited on page 26.

A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker (2007b). Detailed Human Shape and Pose from Images, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. Cited on page 30.

A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker (2007c). Detailed Human Shape and Pose from Images, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. Cited on page 48.

D. Balduzzi, K. Tuyls, J. Pérolat, and T. Graepel (2018). Re-evaluating evaluation, in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Cited on page 187.

A. Banerjee (1994). Initializing Neural Networks using Decision Trees, in *Proceedings of the International Workshop on Computational Learning and Natural Learning Systems*. Cited on page 78.

D. Banica and C. Sminchisescu (2015). Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in RGB-D images, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 139.

A. Bar-Hillel, D. Levi, E. Krupka, and C. Goldberg (2010). Part-Based Feature Synthesis for Human Detection, in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Cited on page 62.

A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Cited on page 193.

H. Barrow and J. M. Tenenbaum (1978). Recovering intrinsic scene characteristics, *Comput. Vis. Syst*, pp. 3–26. Cited on page 190.

B. Barz and J. Denzler (2020). Do We Train on Test Data? Purging CIFAR of Near-Duplicates, *J. Imaging*, vol. 6(6), p. 41. Cited on page 183.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. F. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, Ç. Gülçehre, H. F. Song, A. J. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. R. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu (2018). Relational inductive biases, deep learning, and graph networks, *CoRR*, vol. abs/1806.01261. Cited on page 192.

A. L. Bearman, O. Russakovsky, V. Ferrari, and F. Li (2016). What's the Point: Semantic Segmentation with Point Supervision, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. Cited on page 130.

V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab (2015). Robust Optimization for Deep Regression, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 167.

R. Benenson, M. Mathias, R. Timofte, and L. V. Gool (2012). Pedestrian detection at 100 frames per second, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on pages 15, 16, 67, and 122.

R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool (2013). Seeking the Strongest Rigid Detector, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 16, 62, 64, 66, 67, 69, 78, 81, 86, and 112.

R. Benenson, M. Omran, J. H. Hosang, and B. Schiele (2014). Ten Years of Pedestrian Detection, What Have We Learned?, in *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*. Cited on pages 12, 59, 65, 76, 81, 96, and 113.

R. Benenson, S. Popov, and V. Ferrari (2019). Large-Scale Interactive Object Segmentation With Human Annotators, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 31.

Y. Bengio (2009). Learning Deep Architectures for AI, *Found. Trends Mach. Learn.*, vol. 2(1), pp. 1–127. Cited on page 189.

G. Bertasius, J. Shi, and L. Torresani (2015). DeepEdge: A multi-scale bifurcated deep network for top-down contour detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 141.

N. Bodla, B. Singh, R. Chellappa, and L. S. Davis (2017). Soft-NMS - Improving Object Detection with One Line of Code, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 23.

F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black (2016). Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Cited on pages 29, 32, 36, 45, 48, 161, 173, and 181.

D. Bolya, S. Foley, J. Hays, and J. Hoffman (2020). TIDE: A General Toolbox for Identifying Object Detection Errors, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*. Cited on page 185.

A. Bozkurt, B. Esmaeili, D. H. Brooks, J. G. Dy, and J. van de Meent (2019). Evaluating Combinatorial Generalization in Variational Autoencoders, *CoRR*, vol. abs/1911.04594. Cited on page 192.

R. L. Bras, S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal, and Y. Choi (2020). Adversarial Filters of Dataset Biases, in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Cited on page 186.

M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila (2019). EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41(8), pp. 1844–1861. Cited on pages 7, 10, 13, and 198.

G. Brazil and X. Liu (2019). Pedestrian Detection With Autoregressive Network Phases, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 21.

G. Brazil, X. Yin, and X. Liu (2017). Illuminating Pedestrians via Simultaneous Detection and Segmentation, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 22.

G. J. Brostow, J. Fauqueur, and R. Cipolla (2009). Semantic object classes in video: A high-definition ground truth database, *Pattern Recognit. Lett.*, vol. 30(2), pp. 88–97. Cited on pages 122, 123, 125, 126, 133, and 134.

C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner (2019). MONet: Unsupervised Scene Decomposition and Representation, *CoRR*, vol. abs/1901.11390. Cited on page 192.

W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki (2015). Scene labeling with LSTM recurrent neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 130.

Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos (2016). A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*. Cited on page 20.

Z. Cai, M. J. Saberian, and N. Vasconcelos (2015). Learning Complexity-Aware Cascades for Deep Pedestrian Detection, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 97.

J. C. Caicedo and S. Lazebnik (2015). Active Object Localization with Deep Reinforcement Learning, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 15.

J. F. Canny (1986). A Computational Approach to Edge Detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8(6), pp. 679–698. Cited on pages 141, 142, and 146.

C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang (2015). Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 142.

N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin (2019). On Evaluating Adversarial Robustness, *CoRR*, vol. abs/1902.06705. Cited on page 187.

O. Chapelle and M. Wu (2010). Gradient descent optimization of smoothed information retrieval metrics, *Inf. Retr.*, vol. 13(3), pp. 216–235. Cited on page 41.

C. Chen and D. Ramanan (2017). 3D Human Pose Estimation = 2D Pose Estimation + Matching, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 37 and 47.

C. Chen, A. Tyagi, A. Agrawal, D. Drover, R. MV, S. Stojanov, and J. M. Rehg (2019). Unsupervised 3D Pose Estimation With Geometric Self-Supervision, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 46.

G. Chen, Y. Ding, J. Xiao, and T. X. Han (2013). Detection Evolution with Multi-order Contextual Co-occurrence, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 62 and 65.

L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille (2015a). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on pages 17, 20, 130, 131, 132, and 169.

L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam (2018a). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*. Cited on page 180.

T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud (2018b). Neural Ordinary Differential Equations, in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Cited on page 195.

W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen (2016). Synthesizing Training Images for Boosting Human 3D Pose Estimation, in *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*. Cited on page 34.

Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool (2018c). Domain Adaptive Faster R-CNN for Object Detection in the Wild, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 189.

Y. Chen, X. Liu, and M. Yang (2015b). Multi-instance object segmentation with occlusion handling, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 135.

Y. Chen and T. Pock (2017). Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39(6), pp. 1256–1272. Cited on page 196.

Z. Chen, O. Lam, A. Jacobson, and M. Milford (2014). Convolutional Neural Network-based Place Recognition, *CoRR*, vol. abs/1411.1509. Cited on page 75.

M. Cheng, V. A. Prisacariu, S. Zheng, P. H. S. Torr, and C. Rother (2015). DenseCut: Densely Connected CRFs for Realtime GrabCut, *Comput. Graph. Forum*, vol. 34(7), pp. 193–201. Cited on page 145.

Y. Cheng, B. Yang, B. Wang, Y. Wending, and R. T. Tan (2019). Occlusion-Aware Networks for 3D Human Pose Estimation in Video, in *2019 IEEE/CVF International*

Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. Cited on page 44.

H. Choi, G. Moon, and K. M. Lee (2020). Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*. Cited on page 50.

V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black (2020). Monocular Expressive Body Regression Through Body-Driven Attention, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X*. Cited on page 28.

K. J. Cios and L. Ning (1992). A machine learning method for generation of a neural network architecture: a continuous ID3 algorithm, *IEEE Trans. Neural Networks*, vol. 3(2), pp. 280–291. Cited on page 78.

CMU Graphics Lab Motion Capture Database. `http://mocap.cs.cmu.edu`. Cited on pages 31 and 34.

COCO Analysis Toolkit. `http://cocodataset.org/#detection-eval`. Cited on page 185.

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on pages 12 and 121.

A. D. Costea and S. Nedevschi (2014). Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 62 and 68.

A. D. Costea and S. Nedevschi (2016). Semantic Channels for Fast Pedestrian Detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 22.

F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas (2007). The blur effect: perception and estimation with a new no-reference perceptual blur metric, in *Human Vision and Electronic Imaging XII, San Jose, CA, USA, January 29 - February 1, 2007*. Cited on page 103.

R. Dabral, N. B. Gundavarapu, R. Mitra, A. Sharma, G. Ramakrishnan, and A. Jain (2019). Multi-Person 3D Human Pose Estimation from Monocular Images, in *2019 International Conference on 3D Vision, 3DV 2019, Québec City, QC, Canada, September 16-19, 2019*. Cited on page 53.

R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain (2018). Learning 3D Human Pose from Structure and Motion, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*. Cited on page 44.

J. Dai, K. He, and J. Sun (2015). Convolutional feature masking for joint object and stuff segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 130 and 135.

J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei (2017). Deformable Convolutional Networks, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 196.

N. Dalal and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. Cited on pages 10, 60, 61, 62, 64, 68, 69, and 81.

M. Dantone, J. Gall, C. Leistner, and L. V. Gool (2014). Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36(11), pp. 2131–2143. Cited on page 32.

G. Desjardins, A. C. Courville, and Y. Bengio (2012). Disentangling Factors of Variation via Generative Entangling, *CoRR*, vol. abs/1210.5474. Cited on page 192.

J. Devaranjan, A. Kar, and S. Fidler (2020). Meta-Sim2: Unsupervised Learning of Scene Structure for Synthetic Data Generation, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*. Cited on page 188.

S. K. Divvala, A. A. Efros, and M. Hebert (2012). How Important Are "Deformable Parts" in the Deformable Parts Model?, in *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings, Part III*. Cited on pages 20 and 66.

C. Doersch and A. Zisserman (2019). Sim2real transfer learning for 3D human pose estimation: motion to the rescue, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. Cited on pages 34 and 35.

P. Dollár, R. Appel, S. J. Belongie, and P. Perona (2014). Fast Feature Pyramids for Object Detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36(8), pp. 1532–1545. Cited on pages 16, 62, 86, 87, 96, 112, and 113.

P. Dollár, R. Appel, and W. Kienzle (2012a). Crosstalk Cascades for Frame-Rate Pedestrian Detection, in *Computer Vision - ECCV 2012 - 12th European Conference* *on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*. Cited on page 62.

P. Dollár, S. J. Belongie, and P. Perona (2010). The Fastest Pedestrian Detector in the West, in *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*. Cited on page 62.

P. Dollár, Z. Tu, P. Perona, and S. J. Belongie (2009a). Integral Channel Features, in *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*. Cited on pages 11, 16, 62, 64, 68, 69, 77, 81, and 97.

P. Dollár, Z. Tu, H. Tao, and S. J. Belongie (2007). Feature Mining for Image Classification, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. Cited on page 62.

P. Dollár, C. Wojek, B. Schiele, and P. Perona (2009b). Pedestrian detection: A benchmark, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. Cited on pages 10, 59, 60, 61, and 185.

P. Dollár, C. Wojek, B. Schiele, and P. Perona (2012b). Pedestrian Detection: An Evaluation of the State of the Art, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34(4), pp. 743–761. Cited on pages 10, 11, 61, 76, 78, 94, 95, 96, 98, 106, 111, 122, 127, and 190.

P. Dollár and C. L. Zitnick (2015). Fast Edge Detection Using Structured Forests, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37(8), pp. 1558–1570. Cited on pages 140, 141, 142, 144, and 150.

D. Drover, M. V. Rohith, C. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh (2018). Can 3D Pose Be Learned from 2D Projections Alone?, in *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part IV*. Cited on page 45.

A. Effland, E. Kobler, K. Kunisch, and T. Pock (2020). Variational Networks: An Optimal Control Approach to Early Stopping Variational Methods for Image Restoration, *J. Math. Imaging Vis.*, vol. 62(3), pp. 396–416. Cited on pages 196 and 197.

A. M. Elgammal and C. Lee (2004). Inferring 3D Body Pose from Silhouettes Using Activity Manifold Learning, in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*. Cited on page 43.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry (2019). Exploring the Landscape of Spatial Robustness, in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Cited on pages 188, 190, and 191.

M. Enzweiler and D. M. Gavrila (2009). Monocular Pedestrian Detection: Survey and Experiments, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31(12), pp. 2179–2195. Cited on pages 60 and 122.

M. Enzweiler and D. M. Gavrila (2011). A Multilevel Mixture-of-Experts Framework for Pedestrian Classification, *IEEE Trans. Image Process.*, vol. 20(10), pp. 2967–2979. Cited on page 65.

A. Ess, B. Leibe, K. Schindler, and L. V. Gool (2008). A mobile vision system for robust multi-person tracking, in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. Cited on page 60.

A. Ess, B. Leibe, K. Schindler, and L. V. Gool (2009). Robust Multiperson Tracking from a Mobile Platform, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31(10), pp. 1831–1846. Cited on page 65.

M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman (2015). The Pascal Visual Object Classes Challenge: A Retrospective, *Int. J. Comput. Vis.*, vol. 111(1), pp. 98–136. Cited on pages 9, 11, 84, 122, 127, 128, 131, 135, 140, 142, 143, and 185.

M. Fabbri, F. Lanzi, S. Calderara, S. Alletto, and R. Cucchiara (2020). Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 39.

V. Feldman (2020). Does learning require memorization? a short tale about a long tail, in *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*. Cited on page 187.

V. Feldman and C. Zhang (2020). What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Cited on pages 184 and 187.

P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan (2010). Object Detection with Discriminatively Trained Part-Based Models, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32(9), pp. 1627–1645. Cited on pages 16, 19, 62, 66, 77, and 122.

P. F. Felzenszwalb and D. P. Huttenlocher (2004). Efficient Graph-Based Image Segmentation, *Int. J. Comput. Vis.*, vol. 59(2), pp. 167–181. Cited on pages 140, 141, 145, and 146.

P. F. Felzenszwalb and D. P. Huttenlocher (2005). Pictorial Structures for Object Recognition, *Int. J. Comput. Vis.*, vol. 61(1), pp. 55–79. Cited on page 39.

P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan (2008). A discriminatively trained, multiscale, deformable part model, in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. Cited on pages 61, 62, and 66.

U. Franke, D. Pfeiffer, C. Rabe, C. Knöppel, M. Enzweiler, F. Stein, and R. G. Herrtwich (2013). Making Bertha See, in *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*. Cited on page 122.

P. T. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Mühlfellner, S. Wonneberger, J. Timpner, S. Rottmann, B. Li, B. Schmidt, T. Nguyen, E. Cardarelli, S. Cattani, S. Bruning, S. Horstmann, M. Stellmacher, H. Mielenz, K. Köser, M. Beermann, C. Hane, L. Heng, G. H. Lee, F. Fraundorfer, R. Iser, R. Triebel, I. Posner, P. Newman, L. C. Wolf, M. Pollefeys, S. Brosig, J. Effertz, C. Pradalier, and R. Siegwart (2013). Toward automated driving in cities using close-to-market sensors: An overview of the V-Charge Project, in *2013 IEEE Intelligent Vehicles Symposium (IV), Gold Coast City, Australia, June 23-26, 2013*. Cited on page 122.

V. Gabeur, J. Franco, X. Martin, C. Schmid, and G. Rogez (2019). Moulding Humans: Non-Parametric 3D Human Shape Estimation From Single Images, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 42.

F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele (2013). A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on page 139.

J. Gall, B. Rosenhahn, T. Brox, and H. Seidel (2010). Optimization and Filtering for Human Motion Capture, *Int. J. Comput. Vis.*, vol. 87(1-2), pp. 75–92. Cited on page 48.

Y. Ganin and V. S. Lempitsky (2014). Nˆ4 -Fields: Neural Network Nearest Neighbor Fields for Image Transforms, in *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II*. Cited on page 141.

D. Gavrila and L. S. Davis (1996). 3-D model-based tracking of humans in action: a multi-view approach, in *1996 Conference on Computer Vision and Pattern Recognition (CVPR '96), June 18-20, 1996 San Francisco, CA, USA*. Cited on page 29.

D. M. Gavrila (2007). Looking at people, in *Fourth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, 5-7 September, 2007, Queen Mary, University of London, London, United Kingdom*. Cited on page 2.

D. M. Gavrila and S. Munder (2007). Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle, *Int. J. Comput. Vis.*, vol. 73(1), pp. 41–59. Cited on page 15.

A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun (2014). 3D Traffic Scene Understanding From Movable Platforms, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36(5), pp. 1012–1025. Cited on page 122.

A. Geiger, P. Lenz, C. Stiller, and R. Urtasun (2013). Vision meets robotics: The KITTI dataset, *I. J. Robotics Res.*, vol. 32(11), pp. 1231–1237. Cited on pages 122, 123, 125, and 126.

A. Geiger, P. Lenz, and R. Urtasun (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on pages 11, 12, 60, 78, 87, 95, and 185.

R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann (2020). Shortcut Learning in Deep Neural Networks, *Nature Machine Intelligence*, vol. 2(11), pp. 665–673. Cited on page 180.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. Cited on page 190.

S. Geman and D. E. McClure (1987). Statistical Methods for Tomographic Image Reconstruction, *Bulletin of the International Statistical Institute*, vol. 52(4), pp. 5–21. Cited on page 167.

G. Georgakis, R. Li, S. Karanam, T. Chen, J. Kosecká, and Z. Wu (2020). Hierarchical Kinematic Human Mesh Recovery, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*. Cited on page 49.

M. F. Ghezelghieh, R. Kasturi, and S. Sarkar (2016). Learning Camera Viewpoint Using CNN to Improve 3D Body Pose Estimation, in *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*. Cited on pages 34 and 44.

R. B. Girshick (2015). Fast R-CNN, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on pages 113, 135, 140, 145, and 149.

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 15, 67, 75, 76, 77, 84, 95, 97, 113, and 135.

A. González, Z. Fang, Y. S. Salas, J. Serrat, D. Vázquez, J. Xu, and A. M. López (2016). Pedestrian Detection at Day/Night Time with Visible and FIR Cameras: A Comparison, *Sensors*, vol. 16(6), p. 820. Cited on page 7.

A. Gonzalez-Garcia, A. Vezhnevets, and V. Ferrari (2015). An active search strategy for efficient object class detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 15.

I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio (2013). Maxout Networks, in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Cited on page 197.

S. Gould, R. Hartley, and D. Campbell (2019). Deep Declarative Networks: A New Hope, *CoRR*, vol. abs/1909.04866. Cited on page 195.

K. Grauman, G. Shakhnarovich, and T. Darrell (2003). Inferring 3D Structure with a Statistical Image-Based Shape Model, in *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*. Cited on page 34.

K. Greff, R. L. Kaufman, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner (2019). Multi-Object Representation Learning with Iterative Variational Inference, in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Cited on page 192.

C. Gu, J. J. Lim, P. Arbelaez, and J. Malik (2009). Recognition using regions, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. Cited on page 15.

P. Guan, A. Weiss, A. O. Balan, and M. J. Black (2009). Estimating human shape and pose from a single image, in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. Cited on page 48.

R. A. Güler and I. Kokkinos (2019). HoloPose: Holistic 3D Human Reconstruction In-The-Wild, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 47, 49, and 52.

R. A. Güler, N. Neverova, and I. Kokkinos (2018). DensePose: Dense Human Pose Estimation in the Wild, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 42.

F. Güney and A. Geiger (2015). Displets: Resolving stereo ambiguities using object knowledge, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 127.

A. Gupta, P. Dollár, and R. B. Girshick (2019). LVIS: A Dataset for Large Vocabulary Instance Segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 184.

I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt (2019). In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 36.

S. Hallman and C. C. Fowlkes (2015). Oriented edge forests for boundary detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 141 and 144.

B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik (2011). Semantic contours from inverse detectors, in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*. Cited on pages 140, 141, 142, 143, 144, 155, and 156.

B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik (2014a). Simultaneous Detection and Segmentation, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*. Cited on pages 134 and 135.

B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik (2015). Hypercolumns for object segmentation and fine-grained localization, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 135.

B. Hariharan, C. L. Zitnick, and P. Dollár (2014b). Detecting Objects Using Deformation Dictionaries, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 66.

N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H. Seidel (2009). A Statistical Model of Human Pose and Body Shape, *Comput. Graph. Forum*, vol. 28(2), pp. 337–346. Cited on page 29.

M. Hassan, V. Choutas, D. Tzionas, and M. Black (2019). Resolving 3D Human Pose Ambiguities With 3D Scene Constraints, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 26.

M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black (2021). Populating 3D Scenes by Learning Human-Scene Interaction, pp. 14708–14718. Cited on pages 35 and 188.

H. Hattori, V. N. Boddeti, K. M. Kitani, and T. Kanade (2015). Learning scene-specific pedestrian detectors without real data, in *IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. Cited on page 133.

H. He and B. Upcroft (2013). Nonparametric semantic segmentation for 3D street scenes, in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, Japan, November 3-7, 2013*. Cited on page 127.

K. He, G. Gkioxari, P. Dollár, and R. B. Girshick (2017). Mask R-CNN, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 53 and 196.

K. He, X. Zhang, S. Ren, and J. Sun (2014). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*. Cited on page 76.

K. He, X. Zhang, S. Ren, and J. Sun (2016). Deep Residual Learning for Image Recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on pages 17, 167, 169, and 194.

X. He and S. Gould (2014). An Exemplar-Based CRF for Multi-instance Object Segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 135.

P. Henderson and V. Ferrari (2018). Learning to Generate and Reconstruct 3D Meshes with only 2D Supervision, in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. Cited on page 50.

D. Hendrycks and T. G. Dietterich (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. Cited on pages 189 and 196.

J. Hernández-Orallo (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, *Artif. Intell. Rev.*, vol. 48(3), pp. 397–447. Cited on page 191.

I. Higgins, D. Amos, D. Pfau, S. Racanière, L. Matthey, D. J. Rezende, and A. Lerchner (2018). Towards a Definition of Disentangled Representations, *CoRR*, vol. abs/1812.02230. Cited on page 192.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. Cited on page 192.

D. C. Hogg (1983). Model-based vision: a program to see a walking person, *Image Vis. Comput.*, vol. 1(1), pp. 5–20. Cited on pages 37 and 48.

D. Hoiem, Y. Chodpathumwan, and Q. Dai (2012). Diagnosing Error in Object Detectors, in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*. Cited on page 185.

D. Hoiem, J. Hays, J. Xiao, and A. Khosla (2015). Guest Editorial: Scene Understanding, *Int. J. Comput. Vis.*, vol. 112(2), pp. 131–132. Cited on page 121.

G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie (2018). The INaturalist Species Classification and Detection Dataset, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 184.

J. H. Hosang, R. Benenson, P. Dollár, and B. Schiele (2016). What Makes for Effective Detection Proposals?, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38(4), pp. 814–830. Cited on pages 15, 77, 135, and 139.

J. H. Hosang, R. Benenson, and B. Schiele (2017). Learning Non-maximum Suppression, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 23.

J. H. Hosang, M. Omran, R. Benenson, and B. Schiele (2015). Taking a deeper look at pedestrians, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 75 and 86.

X. Hou, A. L. Yuille, and C. Koch (2013). Boundary Detection Benchmarking: Beyond F-Measures, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on page 144.

S. Huang and D. Ramanan (2017). Expecting the Unexpected: Training Detectors for Unusual Pedestrians with Adversarial Imposters, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 13.

X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang (2018). The ApolloScape Dataset for Autonomous Driving, in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 137.

X. Huang, Z. Ge, Z. Jie, and O. Yoshie (2020). NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 23.

D. Hupkes, V. Dankers, M. Mul, and E. Bruni (2020). Compositionality Decomposed: How do Neural Networks Generalise?, *J. Artif. Intell. Res.*, vol. 67, pp. 757–795. Cited on page 193.

S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon (2015). Multispectral pedestrian detection: Benchmark dataset and baseline, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 7, 10, and 13.

E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele (2016). DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Cited on pages 41 and 167.

C. Ionescu, L. Bo, and C. Sminchisescu (2009). Structural SVM for visual localization and continuous state estimation, in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. Cited on page 34.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36(7), pp. 1325–1339. Cited on pages 30, 31, 34, 51, 166, 179, 183, and 184.

P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson (2014). Crisp Boundary Detection Using Pointwise Mutual Information, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*. Cited on page 141.

I. Ivanova and M. Kubat (1995). Initialization of neural networks by means of decision trees, *Knowl. Based Syst.*, vol. 8(6), pp. 333–344. Cited on page 78.

M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu (2015). Spatial Transformer Networks, in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Cited on page 196.

E. Jahangiri and A. L. Yuille (2017). Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections, in *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 45 and 46.

X. Jia, B. D. Brabandere, T. Tuytelaars, and L. V. Gool (2016). Dynamic Filter Networks, in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Cited on page 196.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell (2014). Caffe: Convolutional Architecture for Fast Feature Embedding, in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. Cited on page 79.

B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang (2018). Acquisition of Localization Confidence for Accurate Object Detection, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*. Cited on page 199.

J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick (2017). CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 193.

S. Johnson and M. Everingham (2010). Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, in *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*. Cited on pages 30 and 32.

S. Johnson and M. Everingham (2011). Learning effective human pose estimation from inaccurate annotation, in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. Cited on pages 30, 31, and 32.

H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh (2019). Panoptic Studio: A Massively Multiview System for Social Interaction Capture, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41(1), pp. 190–204. Cited on page 33.

H. Joo, T. Simon, and Y. Sheikh (2018). Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 29 and 50.

A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik (2018). End-to-End Recovery of Human Shape and Pose, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 37, 45, 48, 49, 51, 161, 162, 172, and 173.

A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik (2019). Learning 3D Human Dynamics From Video, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 53.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li (2014). Large-Scale Video Classification with Convolutional Neural Networks, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 75.

I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua (2018). Learning Latent Representations of 3D Human Pose with Deep Neural Networks, *Int. J. Comput. Vis.*, vol. 126(12), pp. 1326–1341. Cited on page 43.

H. Kato, D. Beker, M. Morariu, T. Ando, T. Matsuoka, W. Kehl, and A. Gaidon (2020). Differentiable Rendering: A Survey, *CoRR*, vol. abs/2006.12057. Cited on page 188.

W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman (2017). The Kinetics Human Action Video Dataset, *CoRR*, vol. abs/1705.06950. Cited on pages 33 and 35.

C. G. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila (2011). The Benefits of Dense Stereo for Pedestrian Detection, *IEEE Trans. Intell. Transp. Syst.*, vol. 12(4), pp. 1096–1106. Cited on page 64.

C. G. Keller, D. F. Llorca, and D. M. Gavrila (2009). Dense Stereo-Based ROI Generation for Pedestrian Detection, in *Pattern Recognition, 31st DAGM Symposium, Jena, Germany, September 9-11, 2009. Proceedings*. Cited on pages 15 and 60.

D. Keysers, N. Schärli, N. Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, T. Tihon, D. Tsarkov, X. Wang, M. van Zee, and O. Bousquet (2020). Measuring Compositional Generalization: A Comprehensive Method on Realistic Data, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. Cited on page 193.

A. Khoreva, R. Benenson, M. Omran, M. Hein, and B. Schiele (2016). Weakly Supervised Object Boundaries, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 139.

S. Kiciroglu, H. Rhodin, S. N. Sinha, M. Salzmann, and P. Fua (2020). ActiveMoCap: Optimized Viewpoint Selection for Active Human Motion Capture, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 44.

H. Kim and A. Mnih (2018). Disentangling by Factorising, in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Cited on page 192.

W. Kim, M. S. Ramanagopal, C. Barto, M. Yu, K. Rosaen, N. Goumas, R. Vasudevan, and M. Johnson-Roberson (2019). PedX: Benchmark Dataset for Metric 3-D Pose Estimation of Pedestrians in Complex Urban Intersections, *IEEE Robotics Autom. Lett.*, vol. 4(2), pp. 1940–1947. Cited on page 33.

S. Kinauer, R. A. Güler, S. Chandra, and I. Kokkinos (2017). Structured Output Prediction and Learning for Deep Monocular 3D Human Pose Estimation, in *Energy Minimization Methods in Computer Vision and Pattern Recognition - 11th International Conference, EMMCVPR 2017, Venice, Italy, October 30 - November 1, 2017, Revised Selected Papers*. Cited on pages 39 and 40.

D. P. Kingma and J. Ba (2015). Adam: A Method for Stochastic Optimization, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on page 167.

E. Kobler, A. Effland, K. Kunisch, and T. Pock (2020). Total Deep Variation for Linear Inverse Problems, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 196.

E. Kobler, T. Klatzer, K. Hammernik, and T. Pock (2017). Variational Networks: Connecting Variational Methods and Deep Learning, in *Pattern Recognition - 39th German Conference, GCPR 2017, Basel, Switzerland, September 12-15, 2017, Proceedings*. Cited on page 196.

M. Kocabas, N. Athanasiou, and M. J. Black (2020). VIBE: Video Inference for Human Body Pose and Shape Estimation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 53.

P. W. Koh and P. Liang (2017). Understanding Black-box Predictions via Influence Functions, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Cited on page 187.

A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby (2020). Big Transfer (BiT): General Visual Representation Learning, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*. Cited on page 183.

N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis (2019a). Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on pages 37 and 51.

N. Kolotouros, G. Pavlakos, and K. Daniilidis (2019b). Convolutional Mesh Regression for Single-Image Human Shape Reconstruction, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 49, 50, and 51.

I. Kostrikov and J. Gall (2014). Depth Sweep Regression Forests for Estimating 3D Human Pose from Images, in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*. Cited on page 39.

K. Krishna, A. Roy, and M. Iyyer (2021). Hurdles to Progress in Long-form Question Answering, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Cited on page 183.

A. Krizhevsky (2009). Learning Multiple Layers of Features from Tiny Images, Technical report, University of Toronto. Cited on page 79.

A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). ImageNet Classification with Deep Convolutional Neural Networks, in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Cited on pages 17, 67, 75, 76, 78, 79, 84, 94, and 122.

L. E. Kruger, C. Wohler, A. Wurz-Wessel, and F. Stein (2004). In-factory calibration of multiocular camera systems, in *Optical Metrology in Production Engineering*. Cited on page 124.

A. Kundu, Y. Li, F. Dellaert, F. Li, and J. M. Rehg (2014). Joint Semantic Segmentation and 3D Reconstruction from Monocular Video, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*. Cited on page 127.

L. Ladicky, J. Shi, and M. Pollefeys (2014). Pulling Things out of Perspective, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 127.

S. Laine, J. Hellsten, T. Karras, Y. Seol, J. Lehtinen, and T. Aila (2020). Modular primitives for high-performance differentiable rendering, *ACM Trans. Graph.*, vol. 39(6), pp. 194:1–194:14. Cited on page 188.

B. M. Lake and M. Baroni (2018). Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks, in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Cited on page 193.

J. P. Lalor, H. Wu, and H. Yu (2016). Building an Evaluation Scale using Item Response Theory, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. Cited on page 187.

S. Lapuschkin, A. Binder, G. Montavon, K. Müller, and W. Samek (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 185.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller (2019). Unmasking Clever Hans predictors and assessing what machines really learn, *Nature Communications*, vol. 10(1), p. 1096. Cited on page 180.

C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler (2017). Unite the People: Closing the Loop Between 3D and 2D Human Representations, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 30, 32, 33, 42, 51, 161, 163, 165, 166, 170, 173, and 181.

H. Law and J. Deng (2018). CornerNet: Detecting Objects as Paired Keypoints, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*. Cited on page 22.

Y. LeCun, Y. Bengio, and G. E. Hinton (2015). Deep learning, *Nat.*, vol. 521(7553), pp. 436–444. Cited on page 122.

D. Lee, G. Cha, M. Yang, and S. Oh (2016). Individualness and Determinantal Point Processes for Pedestrian Detection, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Cited on page 23.

H. Lee and Z. Chen (1985). Determination of 3D human body postures from a single view, *Comput. Vis. Graph. Image Process.*, vol. 30(2), pp. 148–168. Cited on pages 37 and 48.

A. M. Lehrmann, P. V. Gehler, and S. Nowozin (2013). A Non-parametric Bayesian Network Prior of Human Pose, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 45 and 191.

B. Leibe, N. Cornelis, K. Cornelis, and L. V. Gool (2007). Dynamic 3D Scene Analysis from a Moving Vehicle, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. Cited on page 122.

B. Leibe, A. Leonardis, and B. Schiele (2008). Robust Object Detection with Interleaved Categorization and Segmentation, *Int. J. Comput. Vis.*, vol. 77(1-3), pp. 259–289. Cited on page 135.

D. Levi, S. Silberstein, and A. Bar-Hillel (2013). Fast Multiple-Part Based Object Detection Using KD-Ferns, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on page 62.

S. Levine, C. Finn, T. Darrell, and P. Abbeel (2016). End-to-End Training of Deep Visuomotor Policies, *J. Mach. Learn. Res.*, vol. 17, pp. 39:1–39:40. Cited on page 41.

J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan (2018). Scale-Aware Fast R-CNN for Pedestrian Detection, *IEEE Trans. Multimedia*, vol. 20(4), pp. 985–996. Cited on pages 19, 20, and 97.

S. Li and A. B. Chan (2014). 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network, in *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II*. Cited on pages 35, 38, and 40.

S. Li, W. Zhang, and A. B. Chan (2015). Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 43.

Y. Li, M. Paluri, J. M. Rehg, and P. Dollár (2016). Unsupervised Learning of Edges, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 141.

Y. Li and N. Vasconcelos (2019). REPAIR: Removing Representation Bias by Dataset Resampling, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 186.

Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman (2019). Learning the Depths of Moving People by Watching Frozen People, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 2.

J. Liang and M. C. Lin (2019). Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 34.

J. J. Lim, C. L. Zitnick, and P. Dollár (2013). Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on page 68.

C. Lin, J. Lu, G. Wang, and J. Zhou (2018). Graininess-Aware Deep Feature Learning for Pedestrian Detection, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*. Cited on page 22.

G. Lin, A. Milan, C. Shen, and I. D. Reid (2017a). RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 167.

G. Lin, C. Shen, A. van den Hengel, and I. D. Reid (2016). Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on pages 130, 131, 132, and 133.

T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie (2017b). Feature Pyramid Networks for Object Detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 17 and 194.

T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár (2017c). Focal Loss for Dense Object Detection, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 16.

T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). Microsoft COCO: Common Objects in Context, in *Computer Vision -*

ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Cited on pages 11, 122, 127, 130, 134, 137, 142, 143, 183, 184, and 190.

Z. Lin and L. S. Davis (2008). A Pose-Invariant Descriptor for Human Detection and Segmentation, in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV*. Cited on page 62.

D. Linsley, A. K. Ashok, L. N. Govindarajan, R. Liu, and T. Serre (2020). Stable and expressive recurrent vision models, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Cited on pages 195 and 196.

D. Linsley, J. Kim, V. Veerabadran, C. Windolf, and T. Serre (2018). Learning long-range spatial dependencies with horizontal gated recurrent units, in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Cited on page 196.

D. Liu, Z. Zhao, X. Wang, Y. Hu, L. Zhang, and T. S. Huang (2019a). Improving 3D Human Pose Estimation Via 3D Part Affinity Fields, in *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. Cited on page 41.

H. D. Liu, M. Tao, C. Li, D. Nowrouzezahrai, and A. Jacobson (2019b). Beyond Pixel Norm-Balls: Parametric Adversaries using an Analytically Differentiable Renderer, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. Cited on page 188.

L. Liu, W. Ouyang, X. Wang, P. W. Fieguth, J. Chen, X. Liu, and M. Pietikäinen (2020). Deep Learning for Generic Object Detection: A Survey, *Int. J. Comput. Vis.*, vol. 128(2), pp. 261–318. Cited on page 13.

M. Liu, T. Breuel, and J. Kautz (2017). Unsupervised Image-to-Image Translation Networks, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Cited on page 137.

S. Liu, D. Huang, and Y. Wang (2019c). Adaptive NMS: Refining Pedestrian Detection in a Crowd, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 23.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg (2016). SSD: Single Shot MultiBox Detector, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. Cited on pages 17 and 21.

W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen (2018). Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*. Cited on page 21.

W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu (2019d). High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 18 and 21.

W. Liu, A. Rabinovich, and A. C. Berg (2015a). ParseNet: Looking Wider to See Better, *CoRR*, vol. abs/1506.04579. Cited on page 130.

Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang (2015b). Semantic Image Segmentation via Deep Parsing Network, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on pages 130 and 132.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem (2019). Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Cited on page 192.

F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf (2020). Object-Centric Learning with Slot Attention, in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Cited on page 192.

J. Long, E. Shelhamer, and T. Darrell (2015). Fully convolutional networks for semantic segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 121, 122, 125, 129, 130, and 131.

J. Long, N. Zhang, and T. Darrell (2014). Do Convnets Learn Correspondence?, in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Cited on page 75.

M. Loper, N. Mahmood, and M. J. Black (2014). MoSh: motion and shape capture from sparse markers, *ACM Trans. Graph.*, vol. 33(6), pp. 220:1–220:13. Cited on pages 31, 34, and 166.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black (2015). SMPL: a skinned multi-person linear model, *ACM Trans. Graph.*, vol. 34(6), pp. 248:1–248:16. Cited on pages 28, 29, 42, 48, 162, and 163.

M. M. Loper and M. J. Black (2014). OpenDR: An Approximate Differentiable Renderer, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*. Cited on page 50.

Y. Lu, T. Javidi, and S. Lazebnik (2016). Adaptive Object Detection Using Adjacency and Zoom Prediction, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 15.

Z.-L. Lu and B. Dosher (2013). *Visual Psychophysics: From Laboratory to Theory*, The MIT Press. Cited on page 188.

C. Luo, X. Chu, and A. L. Yuille (2018). OriNet: A Fully Convolutional Network for 3D Human Pose Estimation, in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. Cited on page 41.

P. Luo, Y. Tian, X. Wang, and X. Tang (2014). Switchable Deep Network for Pedestrian Detection, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 19, 62, 67, 68, 79, and 82.

Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun (2020). Where, What, Whether: Multi-Modal Learning Meets Pedestrian Detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 22.

D. C. Luvizon, D. Picard, and H. Tabia (2018). 2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 26, 41, and 44.

N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black (2019). AMASS: Archive of Motion Capture As Surface Shapes, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 31.

M. Maire, S. X. Yu, and P. Perona (2011). Object detection and segmentation from joint embedding of parts and pixels, in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*. Cited on page 135.

S. Maji, A. C. Berg, and J. Malik (2008). Classification using intersection kernel support vector machines is efficient, in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. Cited on pages 62 and 64.

J. Mao, T. Xiao, Y. Jiang, and Z. Cao (2017). What Can Help Pedestrian Detection?, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 22.

J. Marín, D. Vázquez, A. M. López, J. Amores, and B. Leibe (2013). Random Forests of Local Experts for Pedestrian Detection, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 62, 65, and 68.

D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik (2001). A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics, in *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 2*. Cited on page 142.

G. H. Martinez, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh (2019). Single-Network Whole-Body Pose Estimation, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 28.

J. Martinez, R. Hossain, J. Romero, and J. J. Little (2017). A Simple Yet Effective Baseline for 3d Human Pose Estimation, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 36, 38, 39, 46, 50, and 53.

F. Martínez-Plumed, R. B. C. Prudêncio, A. M. Usó, and J. Hernández-Orallo (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level, *Artif. Intell.*, vol. 271, pp. 18–42. Cited on page 187.

S. Mathe, A. Pirinen, and C. Sminchisescu (2016). Reinforcement Learning for Visual Object Detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 15.

M. Mathias, R. Benenson, R. Timofte, and L. V. Gool (2013). Handling Occlusions with Franken-Classifiers, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on page 62.

D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt (2017a). Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision, in *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*. Cited on pages 30, 32, 38, 40, and 51.

D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt (2020). XNect: real-time multi-person 3D motion capture with a single RGB camera, *ACM Trans. Graph.*, vol. 39(4), p. 82. Cited on page 54.

D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt (2018). Single-Shot Multi-person 3D Pose Estimation from Monocular RGB, in *2018 International Conference on 3D Vision, 3DV 2018, Verona, Italy, September 5-8, 2018*. Cited on pages 30, 33, and 54.

D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H. Seidel, W. Xu, D. Casas, and C. Theobalt (2017b). VNect: real-time 3D human pose estimation with a single RGB camera, *ACM Trans. Graph.*, vol. 36(4), pp. 44:1–44:14. Cited on pages 40 and 54.

D. N. Metaxas and D. Terzopoulos (1993). Shape and Nonrigid Motion Estimation Through Physics-Based Synthesis, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15(6), pp. 580–591. Cited on page 29.

I. Misra, A. Shrivastava, and M. Hebert (2015). Watch and learn: Semi-supervised learning of object detectors from videos, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 133.

G. Moon, J. Y. Chang, and K. M. Lee (2019). Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 53.

G. Moon and K. M. Lee (2020). I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*. Cited on page 50.

F. Moreno-Noguer (2017). 3D Human Pose Estimation from a Single Image via Distance Matrix Regression, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 36 and 38.

G. Mori and J. Malik (2006). Recovering 3D Human Body Configurations Using Shape Contexts, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28(7), pp. 1052–1062. Cited on page 47.

M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich (2015). Feedforward semantic segmentation with zoom-out features, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 130.

R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. L. Yuille (2014). The Role of Context for Object Detection and Semantic Segmentation in the Wild, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 122 and 130.

P. Mrázek and M. Navara (2003). Selection of Optimal Stopping Time for Nonlinear Diffusion Filtering, *Int. J. Comput. Vis.*, vol. 52(2-3), pp. 189–203. Cited on page 195.

S. Munder and D. M. Gavrila (2006). An Experimental Study on Pedestrian Classification, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28(11), pp. 1863–1868. Cited on page 198.

W. Nam, P. Dollár, and J. H. Han (2014). Local Decorrelation For Improved Pedestrian Detection, in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Cited on pages 12, 68, 69, 76, 77, 87, 96, 97, 108, 113, 177, and 198.

W. Nam, B. Han, and J. H. Han (2011). Improving object localization using macrofeature layout selection, in *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*. Cited on page 62.

G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder (2017). The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 137 and 183.

L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, O. Prokofyeva, R. Thiel, A. Vedaldi, A. Zisserman, and B. Schiele (2018). NightOwls: A Pedestrians at Night Dataset, in *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part I*. Cited on pages 7, 10, and 13.

A. Newell, K. Yang, and J. Deng (2016). Stacked Hourglass Networks for Human Pose Estimation, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*. Cited on pages 39 and 53.

A. Nibali, Z. He, S. Morgan, and L. Prendergast (2018). Numerical Coordinate Regression with Convolutional Neural Networks, *CoRR*, vol. abs/1801.07372. Cited on page 41.

A. Nibali, Z. He, S. Morgan, and L. Prendergast (2019). 3D Human Pose Estimation With 2D Marginal Heatmaps, in *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. Cited on page 41.

M. Nimier-David, S. Speierer, B. Ruiz, and W. Jakob (2020). Radiative backpropagation: an adjoint method for lightning-fast differentiable rendering, *ACM Trans. Graph.*, vol. 39(4), p. 146. Cited on page 188.

J. Noh, S. Lee, B. Kim, and G. Kim (2018). Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 22.

A. Oliva and A. Torralba (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *Int. J. Comput. Vis.*, vol. 42(3), pp. 145–175. Cited on page 121.

M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele (2018). Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation, in *2018 International Conference on 3D Vision, 3DV 2018, Verona, Italy, September 5-8, 2018*. Cited on pages 37, 51, and 161.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic (2015). Is object localization for free? - Weakly-supervised learning with convolutional neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 133 and 142.

A. A. A. Osman, T. Bolkart, and M. J. Black (2020). STAR: Sparse Trained Articulated Human Body Regressor, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*. Cited on page 29.

W. Ouyang and X. Wang (2012). A discriminative deep model for pedestrian detection with occlusion handling, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on pages 19, 62, and 67.

W. Ouyang and X. Wang (2013a). Joint Deep Learning for Pedestrian Detection, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 19, 20, 62, 64, 66, 67, and 82.

W. Ouyang and X. Wang (2013b). Single-Pedestrian Detection Aided by Multi-pedestrian Detection, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 59, 62, 66, 70, and 96.

W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. C. Loy, and X. Tang (2015). DeepID-Net: Deformable deep convolutional neural networks for object detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 76.

W. Ouyang, X. Zeng, and X. Wang (2013). Modeling Mutual Visibility Relationship in Pedestrian Detection, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 19, 62, 66, and 67.

S. Paisitkriangkrai, C. Shen, and A. van den Hengel (2013). Efficient Pedestrian Detection by Directly Optimizing the Partial Area under the ROC Curve, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 62 and 68.

S. Paisitkriangkrai, C. Shen, and A. van den Hengel (2014). Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features, in *Computer Vision -*

ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV. Cited on pages 68, 72, 76, 77, 82, 86, and 87.

G. Papandreou, L. Chen, K. P. Murphy, and A. L. Yuille (2015). Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on pages 130, 132, and 133.

G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf (2018). Learning Independent Causal Mechanisms, in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Cited on page 192.

D. Park, D. Ramanan, and C. C. Fowlkes (2010). Multiresolution Models for Object Detection, in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Cited on pages 62, 63, 65, 66, 67, and 70.

D. Park, C. L. Zitnick, D. Ramanan, and P. Dollár (2013). Exploring Weak Stabilization for Motion Feature Extraction, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 62, 65, and 96.

S. Park, J. Hwang, and N. Kwak (2016). 3D Human Pose Estimation Using Convolutional Neural Networks with 2D Pose Information, in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*. Cited on page 38.

D. Pathak, P. Krähenbühl, and T. Darrell (2015a). Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 130.

D. Pathak, E. Shelhamer, J. Long, and T. Darrell (2015b). Fully Convolutional Multi-Class Multiple Instance Learning, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. Cited on page 130.

A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna (2020). Data and its (dis)contents: A survey of dataset development and use in machine learning research, *CoRR*, vol. abs/2012.05345. Cited on page 182.

G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black (2019a). Expressive Body Capture: 3D Hands, Face, and Body From a Single Image, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 28, 29, 45, and 191.

G. Pavlakos, N. Kolotouros, and K. Daniilidis (2019b). TexturePose: Supervising Human Mesh Estimation With Texture Consistency, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on pages 50 and 52.

G. Pavlakos, X. Zhou, and K. Daniilidis (2018a). Ordinal Depth Supervision for 3D Human Pose Estimation, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 43.

G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis (2017). Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 35, 39, 47, and 50.

G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis (2018b). Learning to Estimate 3D Human Pose and Shape From a Single Color Image, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 37, 48, 49, 51, 162, 165, 172, and 173.

M. Pedersoli, R. Timofte, T. Tuytelaars, and L. V. Gool (2014). Using a Deformation Field Model for Localizing Faces and Facial Points under Weak Supervision, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 66.

X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko (2018). VisDA: A Synthetic-to-Real Benchmark for Visual Domain Adaptation, in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 189.

A. Pentland (2000). Looking at People: Sensing for Ubiquitous and Wearable Computing, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22(1), pp. 107–119. Cited on page 2.

D. Pfeiffer, S. Gehrig, and N. Schneider (2013). Exploiting the Power of Stereo Confidences, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on page 124.

P. H. O. Pinheiro and R. Collobert (2014). Recurrent Convolutional Neural Networks for Scene Labeling, in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Cited on pages 67 and 130.

P. H. O. Pinheiro and R. Collobert (2015). From image-level to pixel-level labeling with Convolutional Networks, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 130 and 142.

R. Plänkers and P. Fua (2001). Articulated Soft Objects for Video-based Body Modeling, in *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01),* *Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 1*. Cited on page 29.

T. Pohlen, A. Hermans, M. Mathias, and B. Leibe (2017). Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 17.

G. Pons-Moll, D. J. Fleet, and B. Rosenhahn (2014). Posebits for Monocular Human Pose Estimation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 43.

G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. W. Fitzgibbon (2015). Metric Regression Forests for Correspondence Estimation, *Int. J. Comput. Vis.*, vol. 113(3), pp. 163–175. Cited on page 42.

J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marqués, and J. Malik (2017). Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39(1), pp. 128–140. Cited on pages 15, 140, 141, 144, and 145.

J. Pont-Tuset and L. V. Gool (2015). Boosting Object Proposals: From Pascal to COCO, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 135.

A. Popa, M. Zanfir, and C. Sminchisescu (2017). Deep Multitask Architecture for Integrated 2D and 3D Human Sensing, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 53.

C. Premebida, J. Carreira, J. Batista, and U. Nunes (2014). Pedestrian detection combining RGB and dense LIDAR data, in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, September 14-18, 2014*. Cited on page 65.

A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari (2012). Learning object class detectors from weakly annotated video, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on page 142.

W. Qiu and A. L. Yuille (2016). UnrealCV: Connecting Computer Vision to Unreal Engine, in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*. Cited on page 188.

I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna (2021). AI and the Everything in the Whole Wide World Benchmark. Cited on page 180.

V. Ramakrishna, T. Kanade, and Y. Sheikh (2012). Reconstructing 3D Human Pose from 2D Image Landmarks, in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*. Cited on pages 47 and 173.

A. Rasouli, I. Kotseruba, and J. K. Tsotsos (2017). Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior, in *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*. Cited on page 8.

A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson (2014). CNN Features Off-the-Shelf: An Astounding Baseline for Recognition, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 75 and 86.

J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi (2016). You Only Look Once: Unified, Real-Time Object Detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 18.

S. Ren, K. He, R. B. Girshick, and J. Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. Cited on pages 16, 17, 18, 20, 47, 49, 135, and 140.

X. Ren and L. Bo (2012). Discriminatively Trained Sparse Code Gradients for Contour Detection, in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. Cited on page 144.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Cited on page 191.

B. RichardWebster, S. E. Anthony, and W. J. Scheirer (2018a). PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41(9), pp. 2280–2286. Cited on page 188.

B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer (2018b). Visual Psychophysics for Making Face Recognition Algorithms More Explainable, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*. Cited on page 188.

S. R. Richter, V. Vineet, S. Roth, and V. Koltun (2016). Playing for Data: Ground Truth from Computer Games, in *Computer Vision - ECCV 2016 - 14th European*

Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II. Cited on page 188.

H. Riemenschneider, A. Bódis-Szomorú, J. Weissenberg, and L. V. Gool (2014). Learning Where to Classify in Multi-view Semantic Segmentation, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Cited on page 126.

H. Riemenschneider, S. Sternig, M. Donoser, P. M. Roth, and H. Bischof (2012). Hough Regions for Joining Instance Localization and Segmentation, in *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III*. Cited on page 135.

M. Roberts and N. Paczan (2020). Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding, *CoRR*, vol. abs/2011.02523. Cited on page 188.

K. M. Robinette and H. A. M. Daanen (1999). The Caesar Project: A 3-D Surface Anthropometry Survey, in *2nd International Conference on 3D Digital Imaging and Modeling (3DIM '99), 4-8 October 1999, Ottawa, Canada*. Cited on pages 31 and 50.

G. Rogez and C. Schmid (2016). MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild, in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. Cited on pages 34 and 35.

G. Rogez, P. Weinzaepfel, and C. Schmid (2020). LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42(5), pp. 1146–1161. Cited on pages 47 and 53.

Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy (2019). Delving Deep Into Hybrid Annotations for 3D Human Recovery in the Wild, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on pages 42, 49, 51, and 52.

G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vázquez, and A. M. L. Peña (2015). Vision-Based Offline-Online Perception Paradigm for Autonomous Driving, in *2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, January 5-9, 2015*. Cited on pages 122, 127, 133, and 134.

A. Rosenfeld, R. S. Zemel, and J. K. Tsotsos (2018). The Elephant in the Room, *CoRR*, vol. abs/1808.03305. Cited on page 190.

S. Roth and M. J. Black (2009). Fields of Experts, *Int. J. Comput. Vis.*, vol. 82(2), pp. 205–229. Cited on pages 195 and 196.

R. Rothe, M. Guillaumin, and L. V. Gool (2014). Non-maximum Suppression for Object Detection by Passing Messages Between Windows, in *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I*. Cited on page 199.

C. Rother, V. Kolmogorov, and A. Blake (2004). "GrabCut": interactive foreground extraction using iterated graph cuts, *ACM Trans. Graph.*, vol. 23(3), pp. 309–314. Cited on page 145.

H. A. Rowley, S. Baluja, and T. Kanade (1995). Human Face Detection in Visual Scenes, in *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, USA, November 27-30, 1995*. Cited on page 15.

L. I. Rudin and S. J. Osher (1994). Total Variation Based Image Restoration with Free Local Constraints, in *Proceedings 1994 International Conference on Image Processing, Austin, Texas, USA, November 13-16, 1994*. Cited on page 196.

N. Rueegg, C. Lassner, M. J. Black, and K. Schindler (2020). Chained Representation Cycling: Learning to Estimate 3D Human Pose and Shape by Cycling Between Representations, in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. Cited on page 52.

E. Rusak, L. Schott, R. S. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel (2020). A Simple Way to Make Neural Networks Robust Against Diverse Image Corruptions, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*. Cited on pages 194 and 196.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li (2015a). ImageNet Large Scale Visual Recognition Challenge, *Int. J. Comput. Vis.*, vol. 115(3), pp. 211–252. Cited on pages 75, 76, 78, 122, 144, 183, and 184.

O. Russakovsky, L. Li, and F. Li (2015b). Best of both worlds: Human-machine collaboration for object annotation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 31.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman (2008). LabelMe: A Database and Web-Based Tool for Image Annotation, *Int. J. Comput. Vis.*, vol. 77(1-3), pp. 157–173. Cited on page 124.

S. Sabour, N. Frosst, and G. E. Hinton (2017). Dynamic Routing Between Capsules, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Cited on page 197.

P. Sabzmeydani and G. Mori (2007). Detecting Pedestrians by Learning Shapelet Features, in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. Cited on page 62.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell (2010). Adapting Visual Category Models to New Domains, in *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Cited on page 189.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang (2020). Distributionally Robust Neural Networks, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. Cited on pages 190 and 191.

S. Saito, T. Simon, J. M. Saragih, and H. Joo (2020). PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 2.

C. Sakaridis, D. Dai, and L. V. Gool (2018). Semantic Foggy Scene Understanding with Synthetic Data, *Int. J. Comput. Vis.*, vol. 126(9), pp. 973–992. Cited on pages 13 and 137.

S. Santurkar, D. Tsipras, and A. Madry (2021). BREEDS: Benchmarks for Subpopulation Shift, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. Cited on page 189.

I. Sárándi, T. Linder, K. O. Arras, and B. Leibe (2021). MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation, *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3(1), pp. 16–30. Cited on pages 40 and 41.

T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth (2013). Efficient Multi-cue Scene Segmentation, in *Pattern Recognition - 35th German Conference, GCPR 2013, Saarbrücken, Germany, September 3-6, 2013. Proceedings.* Cited on pages 122 and 123.

T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth (2014). Stixmantics: A Medium-Level Model for Real-Time Semantic Scene Understanding, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Cited on pages 122 and 126.

J. Schmidhuber (1992). Learning Factorial Codes by Predictability Minimization, *Neural Comput.*, vol. 4(6), pp. 863–879. Cited on page 192.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij (2012). On causal and anticausal learning, in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. Cited on pages 180 and 192.

W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis (2009). Human detection using partial least squares analysis, in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. Cited on page 62.

A. G. Schwing and R. Urtasun (2015). Fully Connected Deep Structured Networks, *CoRR*, vol. abs/1503.02351. Cited on pages 130 and 131.

S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr (2013). Urban 3D semantic modelling using stereo vision, in *2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, May 6-10, 2013*. Cited on pages 127, 133, and 134.

S. Sengupta, P. Sturgess, L. Ladicky, and P. H. S. Torr (2012). Automatic dense visual semantic mapping from street-level imagery, in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*. Cited on page 126.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Cited on pages 16, 67, and 122.

P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun (2013). Pedestrian Detection with Unsupervised Multi-stage Feature Learning, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 16, 19, 62, 64, 67, 77, 80, and 83.

I. K. Sethi and M. Otten (1990). Comparison between entropy net and decision tree classifiers, in *IJCNN 1990, International Joint Conference on Neural Networks, San Diego, CA, USA, June 17-21, 1990*. Cited on page 78.

R. Setiono and W. K. Leow (1999). On mapping decision trees and neural networks, *Knowl. Based Syst.*, vol. 12(3), pp. 95–99. Cited on page 78.

G. Shakhnarovich, P. A. Viola, and T. Darrell (2003). Fast Pose Estimation with Parameter-Sensitive Hashing, in *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*. Cited on pages 34 and 47.

S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun (2018). CrowdHuman: A Benchmark for Detecting Human in a Crowd, *CoRR*, vol. abs/1805.00123. Cited on page 13.

A. Sharma, O. Tuzel, and D. W. Jacobs (2015). Deep hierarchical parsing for semantic segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 130.

S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain (2019). Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 46.

R. Shetty, M. Fritz, and B. Schiele (2020). Towards Automated Testing and Robustification by Semantic Adversarial Data Generation, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*. Cited on page 188.

R. Shetty, B. Schiele, and M. Fritz (2019). Not Using the Car to See the Sidewalk - Quantifying and Controlling the Effects of Context in Classification and Segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 190.

H. Sidenbladh and M. J. Black (2001). Learning Image Statistics for Bayesian Tracking, in *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 2*. Cited on page 29.

L. Sigal, A. O. Balan, and M. J. Black (2010). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion, *Int. J. Comput. Vis.*, vol. 87(1-2), pp. 4–27. Cited on pages 30, 31, and 166.

L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard (2004). Tracking Loose-Limbed People, in *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*. Cited on page 29.

E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer (2012). Single image 3D human pose estimation from noisy observations, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on page 46.

K. Simonyan and A. Zisserman (2014). Two-Stream Convolutional Networks for Action Recognition in Videos, in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Cited on page 75.

K. Simonyan and A. Zisserman (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on pages 17, 76, 131, 144, 149, 169, and 194.

C. Sminchisescu and A. D. Jepson (2004). Generative modeling for continuous non-linearly embedded visual inference, in *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*. Cited on page 43.

C. Sminchisescu, A. Kanaujia, Z. Li, and D. N. Metaxas (2005). Discriminative Density Propagation for 3D Human Motion Estimation, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. Cited on page 34.

C. Sminchisescu and B. Triggs (2003). Estimating Articulated Human Motion With Covariance Scaled Sampling, *Int. J. Robotics Res.*, vol. 22(6), pp. 371–392. Cited on pages 37 and 48.

K. Sohn, H. Lee, and X. Yan (2015). Learning Structured Output Representation using Deep Conditional Generative Models, in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada.* Cited on page 46.

S. Song, S. P. Lichtenberg, and J. Xiao (2015). SUN RGB-D: A RGB-D scene understanding benchmark suite, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 126.

T. Song, L. Sun, D. Xie, H. Sun, and S. Pu (2018). Small-Scale Pedestrian Detection Based on Topological Line Localization and Temporal Feature Aggregation, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*. Cited on page 21.

C. J. Spoerer, T. C. Kietzmann, J. Mehrer, I. Charest, and N. Kriegeskorte (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision, *PLoS Computational Biology*, vol. 16(10), pp. e1008215–e1008215. Cited on page 197.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, vol. 15(1), pp. 1929–1958. Cited on page 49.

H. Su, J. Deng, and L. Fei-Fei (2012). Crowdsourcing Annotations for Visual Object Detection, in *The 4th Human Computation Workshop, HCOMP@AAAI 2012, Toronto, Ontario, Canada, July 23, 2012*. Cited on page 31.

P. Sudowe and B. Leibe (2011). Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video, in *Computer Vision Systems - 8th International Conference, ICVS 2011, Sophia Antipolis, France, September 20-22, 2011. Proceedings.* Cited on page 15.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta (2017a). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 183.

X. Sun, J. Shang, S. Liang, and Y. Wei (2017b). Compositional Human Pose Regression, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 40.

X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei (2018). Integral Human Pose Regression, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*. Cited on page 41.

Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei (2019). Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 53.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). Going deeper with convolutions, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 75 and 76.

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 83.

M. Tan, R. Pang, and Q. V. Le (2020). EfficientDet: Scalable and Efficient Object Detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on pages 17, 194, and 198.

V. Tan, I. Budvytis, and R. Cipolla (2017). Indirect deep structured learning for 3D human body shape and pose prediction, in *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. Cited on page 51.

M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox (2019). What Do Single-View 3D Reconstruction Networks Learn?, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 47, 183, and 184.

C. J. Taylor (2000). Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. Cited on page 48.

J. Taylor, J. Shotton, T. Sharp, and A. W. Fitzgibbon (2012). The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on page 42.

B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua (2016). Structured Prediction of 3D Human Pose with Deep Neural Networks, in *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. Cited on page 43.

B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua (2017). Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 36 and 40.

Y. Tian, P. Luo, X. Wang, and X. Tang (2015a). Deep Learning Strong Parts for Pedestrian Detection, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on pages 19, 20, and 97.

Y. Tian, P. Luo, X. Wang, and X. Tang (2015b). Pedestrian detection aided by deep learning semantic tasks, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 22, 98, 113, and 115.

Z. Tian, C. Shen, H. Chen, and T. He (2019). FCOS: Fully Convolutional One-Stage Object Detection, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 18.

J. Tighe and S. Lazebnik (2013). Superparsing - Scalable Nonparametric Image Parsing with Superpixels, *Int. J. Comput. Vis.*, vol. 101(2), pp. 329–349. Cited on page 126.

J. Tighe, M. Niethammer, and S. Lazebnik (2015). Scene Parsing with Object Instance Inference Using Regions and Per-exemplar Detectors, *Int. J. Comput. Vis.*, vol. 112(2), pp. 150–171. Cited on pages 121 and 135.

D. Tomè, C. Russell, and L. Agapito (2017). Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 37 and 47.

J. Tompson, A. Jain, Y. LeCun, and C. Bregler (2014). Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation, in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Cited on pages 39 and 75.

A. Torralba and A. A. Efros (2011). Unbiased look at dataset bias, in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. Cited on pages 179 and 185.

A. Toshev and C. Szegedy (2014). DeepPose: Human Pose Estimation via Deep Neural Networks, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 75.

F. Träuble, E. Creager, N. Kilbertus, A. Goyal, F. Locatello, B. Schölkopf, and S. Bauer (2020). Is Independence all you need? On the Generalization of Representations Learned from Correlated Data, *CoRR*, vol. abs/2006.07886. Cited on page 192.

M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. P. Collomosse (2017). Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors, in *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. Cited on page 32.

Z. Tu and X. Bai (2010). Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32(10), pp. 1744–1757. Cited on page 65.

S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths (2021). Are Convolutional Neural Networks or Transformers more like human vision?, *CoRR*, vol. abs/2105.07197. Cited on page 190.

H. Tung, H. Tung, E. Yumer, and K. Fragkiadaki (2017a). Self-supervised Learning of Motion Capture, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. Cited on pages 37, 48, 52, 162, and 173.

H. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki (2017b). Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 45.

J. R. R. Uijlings and V. Ferrari (2015). Situational object boundary detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 142, 143, 149, 150, 155, and 156.

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders (2013). Selective Search for Object Recognition, *Int. J. Comput. Vis.*, vol. 104(2), pp. 154–171. Cited on pages 15, 76, 135, 140, and 145.

B. Uzkent, C. Yeh, and S. Ermon (2020). Efficient Object Detection in Large Images Using Deep Reinforcement Learning, in *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. Cited on page 15.

R. S. van Bergen and N. Kriegeskorte (2020). Going in circles is the way forward: the role of recurrence in visual inference, *Current Opinion in Neurobiology*, vol. 65, pp. 176–193. Cited on page 194.

S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber (2018). Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions, in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. Cited on page 192.

G. Varol, D. Ceylan, B. C. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid (2018). BodyNet: Volumetric Inference of 3D Human Body Shapes, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*. Cited on pages 42, 49, and 50.

G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid (2017). Learning from Synthetic Humans, in *2017 IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. Cited on page 34.

A. Vezhnevets, V. Ferrari, and J. M. Buhmann (2011). Weakly supervised semantic segmentation with a multi-image model, in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*. Cited on page 142.

V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr (2015). Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction, in *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. Cited on page 134.

P. A. Viola and M. J. Jones (2001). Robust Real-Time Face Detection, in *Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 2*. Cited on page 61.

P. A. Viola and M. J. Jones (2004). Robust Real-Time Face Detection, *Int. J. Comput. Vis.*, vol. 57(2), pp. 137–154. Cited on pages 16 and 62.

P. A. Viola, M. J. Jones, and D. Snow (2003). Detecting Pedestrians Using Patterns of Motion and Appearance, in *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*. Cited on pages 61, 69, and 177.

T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll (2018). Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*. Cited on pages 32 and 184.

S. Walk, N. Majer, K. Schindler, and B. Schiele (2010). New features and insights for pedestrian detection, in *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. Cited on pages 62, 65, 68, and 77.

B. Wandt and B. Rosenhahn (2019). RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 46.

C. Wang, W. Ren, K. Huang, and T. Tan (2014a). Weakly Supervised Object Localization with Latent Category Learning, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*. Cited on page 142.

C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao (2014b). Robust Estimation of 3D Human Poses from a Single Image, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 47.

J. Wang, S. Huang, X. Wang, and D. Tao (2019). Not All Parts Are Created Equal: 3D Pose Estimation by Modeling Bi-Directional Dependencies of Body Parts, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 43.

J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao (2020). Deep High-Resolution Representation Learning for Visual Recognition, *to appear in IEEE Trans. Pattern Anal. Mach. Intell.*. Cited on page 17.

S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou (2018a). Pedestrian Detection via Body Part Semantic and Contextual Information With DNN, *IEEE Trans. Multimedia*, vol. 20(11), pp. 3148–3159. Cited on page 22.

T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro (2018b). High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 137.

X. Wang, T. X. Han, and S. Yan (2009). An HOG-LBP human detector with partial occlusion handling, in *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*. Cited on pages 62 and 68.

X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen (2018c). Repulsion Loss: Detecting Pedestrians in a Crowd, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 22.

X. Wang, M. Yang, S. Zhu, and Y. Lin (2013). Regionlets for Generic Object Detection, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 15, 76, and 77.

M. Weber, M. Welling, and P. Perona (2000). Unsupervised Learning of Models for Recognition, in *Computer Vision - ECCV 2000, 6th European Conference on Computer Vision, Dublin, Ireland, June 26 - July 1, 2000, Proceedings, Part I*. Cited on page 15.

Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, J. Feng, Y. Zhao, and S. Yan (2017). STC: A Simple to Complex Framework for Weakly-Supervised Semantic Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39(11), pp. 2314–2320. Cited on page 130.

J. Weickert (1998). *Anisotropic diffusion in image processing*, Teubner. Cited on page 195.

P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez (2020). DOPE: Distillation of Part Experts for Whole-Body 3D Pose Estimation in the Wild, in

Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI. Cited on page 28.

F. A. Wichmann, D. H. J. Janssen, R. Geirhos, G. Aguilar, H. H. Schütt, M. Maertens, and M. Bethge (2017). Methods and measurements to compare men against machines, in *Human Vision and Electronic Imaging 2017, Burlingame, CA, USA, 29 January 2017 - 2 February 2017*. Cited on page 188.

C. Wojek and B. Schiele (2008). A Performance Evaluation of Single and Multi-feature People Detection, in *Pattern Recognition, 30th DAGM Symposium, Munich, Germany, June 10-13, 2008, Proceedings*. Cited on pages 62, 64, and 68.

C. Wojek, S. Walk, and B. Schiele (2009). Multi-cue onboard pedestrian detection, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. Cited on page 60.

M. Wrenninge and J. Unger (2018). Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing, *CoRR*, vol. abs/1810.08705. Cited on page 188.

D. Xiang, H. Joo, and Y. Sheikh (2019). Monocular Total Capture: Posing Face, Body, and Hands in the Wild, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on pages 41 and 50.

K. Y. Xiao, L. Engstrom, A. Ilyas, and A. Madry (2021). Noise or Signal: The Role of Image Backgrounds in Object Recognition, in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. Cited on pages 190 and 191.

J. Xie, M. Kiefel, M. Sun, and A. Geiger (2016). Semantic Instance Annotation of Street Scenes by 3D to 2D Label Transfer, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on pages 123 and 125.

S. Xie and Z. Tu (2017). Holistically-Nested Edge Detection, *Int. J. Comput. Vis.*, vol. 125(1-3), pp. 3–18. Cited on pages 140, 141, 144, and 152.

D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe (2017). Learning Cross-Modal Deep Representations for Robust Pedestrian Detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on page 22.

H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu (2020). GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on pages 29 and 191.

J. Xu, A. G. Schwing, and R. Urtasun (2015). Learning to segment under various forms of weak supervision, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 130 and 142.

P. Xu, F. Davoine, J. Bordes, H. Zhao, and T. Denoeux (2013). Information Fusion on Oversegmented Images: An Application for Urban Scene Understanding, in *Proceedings of the 13. IAPR International Conference on Machine Vision Applications, MVA 2013, Kyoto, Japan, May 20-23, 2013*. Cited on page 127.

Y. Xu, S. Zhu, and T. Tung (2019). DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on pages 42, 49, 50, and 52.

J. Yan, Z. Lei, L. Wen, and S. Z. Li (2014). The Fastest Deformable Part Model for Object Detection, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on page 66.

J. Yan, X. Zhang, Z. Lei, S. Liao, and S. Z. Li (2013). Robust Multi-resolution Pedestrian Detection in Traffic Scenes, in *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. Cited on pages 62, 63, 65, 66, and 67.

W. Yang, W. Ouyang, X. Wang, J. S. J. Ren, H. Li, and X. Wang (2018). 3D Human Pose Estimation in the Wild by Adversarial Learning, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on page 45.

J. Yao, S. Fidler, and R. Urtasun (2012). Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*. Cited on page 135.

H. Yasin, U. Iqbal, B. Krüger, A. Weber, and J. Gall (2016). A Dual-Source Approach for 3D Pose Estimation from a Single Image, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 47.

K. M. Yi, E. Trulls, V. Lepetit, and P. Fua (2016). LIFT: Learned Invariant Feature Transform, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Cited on page 41.

Y. Yoshiyasu, R. Sagawa, K. Ayusawa, and A. Murai (2018). Skeleton Transformer Networks: 3D Human Pose and Skinned Mesh from Single RGB Image, in *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV*. Cited on pages 41 and 47.

F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell (2020). BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on pages 13 and 137.

F. Yu and V. Koltun (2016). Multi-Scale Context Aggregation by Dilated Convolutions, in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Cited on pages 17, 20, 130, 132, and 133.

F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao (2015). LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop, *CoRR*, vol. abs/1506.03365. Cited on page 34.

A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Sukthankar, and C. Sminchisescu (2020). Weakly Supervised 3D Human Pose and Shape Reconstruction with Normalizing Flows, in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*. Cited on pages 37, 48, 50, and 52.

A. Zanfir, E. Marinoiu, and C. Sminchisescu (2018a). Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - The Importance of Multiple Scene Constraints, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 52 and 53.

A. Zanfir, E. Marinoiu, M. Zanfir, A. Popa, and C. Sminchisescu (2018b). Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images, in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. Cited on page 54.

J. Zbontar and Y. LeCun (2015). Computing the stereo matching cost with a convolutional neural network, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 75.

X. Zeng, W. Ouyang, and X. Wang (2013). Multi-stage Contextual Deep Learning for Pedestrian Detection, in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. Cited on pages 19 and 62.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2017a). Understanding deep learning requires rethinking generalization, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. Cited on pages 184, 187, and 198.

H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun (2019). DaNet: Decompose-and-aggregate Network for 3D Human Shape and Pose Estimation, in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. Cited on page 49.

L. Zhang, L. Lin, X. Liang, and K. He (2016a). Is Faster R-CNN Doing Well for Pedestrian Detection?, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Cited on pages 20, 21, and 103.

R. Zhang, S. A. Candra, K. Vetter, and A. Zakhor (2015a). Sensor fusion for semantic segmentation of urban scenes, in *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. Cited on page 127.

S. Zhang, C. Bauckhage, and A. B. Cremers (2014). Informed Haar-Like Features Improve Pedestrian Detection, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. Cited on pages 62, 68, 76, and 77.

S. Zhang, R. Benenson, M. Omran, J. H. Hosang, and B. Schiele (2016b). How Far are We from Solving Pedestrian Detection?, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on page 93.

S. Zhang, R. Benenson, M. Omran, J. H. Hosang, and B. Schiele (2018a). Towards Reaching Human Performance in Pedestrian Detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40(4), pp. 973–986. Cited on page 93.

S. Zhang, R. Benenson, and B. Schiele (2015b). Filtered channel features for pedestrian detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on pages 12, 94, 96, 97, 99, 100, 101, 104, 108, and 113.

S. Zhang, R. Benenson, and B. Schiele (2017b). CityPersons: A Diverse Dataset for Pedestrian Detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. Cited on pages 7, 10, 11, 12, 18, 20, 21, 137, 178, 184, and 198.

S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li (2020a). Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Cited on page 18.

S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li (2018b). Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*. Cited on page 22.

S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo (2020b). WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild, *IEEE Trans. Multimedia*, vol. 22(2), pp. 380–393. Cited on pages 7 and 13.

S. Zhang, J. Yang, and B. Schiele (2018c). Occluded Pedestrian Detection Through Guided Attention in CNNs, in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Cited on pages 20 and 23.

Z. Zhang, A. G. Schwing, S. Fidler, and R. Urtasun (2015c). Monocular Object Instance Segmentation and Depth Ordering with CNNs, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on page 135.

S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr (2015). Conditional Random Fields as Recurrent Neural Networks, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. Cited on pages 130, 131, and 132.

Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu (2019). DeepHuman: 3D Human Reconstruction From a Single Image, in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. Cited on page 50.

B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba (2015a). Object Detectors Emerge in Deep Scene CNNs, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Cited on page 17.

B. Zhou, À. Lapedriza, J. Xiao, A. Torralba, and A. Oliva (2014). Learning Deep Features for Scene Recognition using Places Database, in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Cited on pages 75, 78, 85, and 121.

C. Zhou and J. Yuan (2018). Bi-box Regression for Pedestrian Detection and Occlusion Estimation, in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*. Cited on page 23.

X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei (2017). Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on pages 40, 44, and 45.

X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis (2015b). 3D shape estimation from 2D landmarks: A convex relaxation approach, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 173.

X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei (2016a). Deep Kinematic Pose Regression, in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*. Cited on page 42.

X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis (2016b). Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. Cited on pages 36 and 47.

X. Zhou, J. Zhuo, and P. Krähenbühl (2019a). Bottom-Up Object Detection by Grouping Extreme and Center Points, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 22.

Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li (2019b). On the Continuity of Rotation Representations in Neural Networks, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Cited on page 51.

J. Zhu, T. Park, P. Isola, and A. A. Efros (2017a). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 53.

J. Zhu, T. Park, P. Isola, and A. A. Efros (2017b). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. Cited on page 137.

J. Zhu, Z. Yuan, C. Zhang, W. Chi, Y. Ling, and S. Zhang (2020). Crowded Human Detection via an Anchor-pair Network, in *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. Cited on page 23.

Y. Zhu, R. Urtasun, R. Salakhutdinov, and S. Fidler (2015). segDeepM: Exploiting segmentation and context in deep neural networks for object detection, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. Cited on page 139.

C. L. Zitnick and P. Dollár (2014). Edge Boxes: Locating Object Proposals from Edges, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Cited on pages 77 and 80.

C. L. Zitnick, R. Vedantam, and D. Parikh (2016). Adopting Abstract Images for Semantic Scene Understanding, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38(4), pp. 627–638. Cited on page 188.