Training Object Detectors if Only Large Objects are Labeled

Daniel Pototzky[1]
daniel.pototzky@de.bosch.com
Matthias Kirschner[1]
matthias.kirschner@de.bosch.com
Azhar Sultan[1]
azhar.sultan@de.bosch.com
Lars Schmidt-Thieme[2]
schmidt-thieme@ismll.de

[1] Robert Bosch GmbH, Germany
[2] University of Hildesheim, Germany

Abstract

Conventional methods for object detection typically rely on large, well-annotated datasets, which are in short supply due to the high costs of labeling. In this paper, we propose to label only large, easy-to-spot objects. We argue that these contain more pixels and therefore usually more information about the underlying object class than small ones. At the same time, they are easier to spot and hence cheaper to label. Unfortunately, standard supervised learning algorithms do not learn to detect small objects if only large ones are labeled. Instead, they erroneously take up unlabeled objects as negative examples and their accuracy consequently deteriorates. To address that, we propose PCIS, a novel combination of **P**seudo-labels, output **C**onsistency across scales, and an anchor scale-dependent **I**gnore **S**trategy. In experiments on CityPersons, EuroCityPersons, and MS COCO, we show that our approach outperforms existing pseudo-label generation methods as well as an oracle which ensures that anchors overlapping missing annotations are ignored during training. We demonstrate that using our method it is possible to approach the performance of a fully labeled dataset with only a subset of the labels and also to train detectors on extremely sparsely labeled images, e.g. if only 1 out of 200 objects is annotated.

1 Introduction

In recent years, object detectors have made significant progress in both performance and efficiency. However, conventional detectors require access to fully labeled training data, which is costly to obtain. In order to reduce annotation costs, we suggest labeling only large, easy-to-spot objects. We argue that these contain a greater number of pixels and therefore usually more information about the underlying object class than small ones. By downscaling images, we can generate objects of arbitrary size for training the detector even if only large ones are annotated. At the same time, large objects are easier to spot and can therefore be labeled at a lower cost.

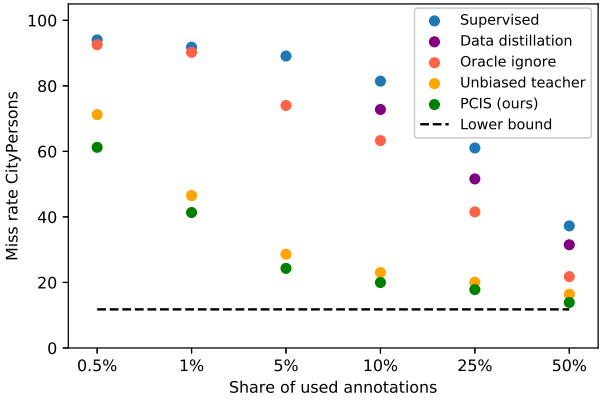

Figure 1: Miss rate on CityPersons [26] by PCIS compared to a supervised baseline, data distillation [14], an oracle ignore, and unbiased teacher [11] if only large objects are labeled. Given the largest 50% of annotations, performance close to the lower bound of using all labels (100% of annotations and ignore labels) can be reached. If only the largest 1% are labeled, our method achieves results that are only slightly inferior to a supervised baseline using 50% of annotations.

Unfortunately, naively following a standard supervised training protocol results in poor performance if only large objects are labeled. Such a detector does not learn to find small objects. Because unlabeled objects are treated as background during training, performance even deteriorates.

Compared to a fully annotated dataset, only labeling large objects (on the same number of images) poses three main challenges. First, the number of bounding boxes per class is much lower. Therefore, the detector is trained on a less diverse set of examples for each class. Second, no examples of small objects are given, meaning that the detector is trained to only detect large ones. Third, missing annotations are treated as background during training, hampering performance.

Previous research on the related problem of dealing with missing annotations has focused on mitigating the third issue by ignoring or downweighting anchors, which are likely to overlap with a missing annotation. Experiments for increasing the number of bounding boxes by pseudo-label based methods did not show any improvement [13, 22].

Contrary to that, we demonstrate that our approach which combines pseudo-labeling, enforcement of consistency in the model output regardless of the input scale and an ignore strategy for anchors which likely contain missing annotations leads to significant gains in performance in experiments on CityPersons [26], EuroCityPersons [1], and MS COCO [8]. Using our technique, it is possible to approach the performance of a fully labeled dataset with only a subset of the labels and to train detectors on extremely few labels (see Fig. 1). We believe these findings to be of immense practical importance. They make it feasible to reduce annotation costs and allow for prototyping given very few labels. Overall, the main contributions of our work can be summarized as follows:

- We suggest labeling only large, easy-to-spot objects given a limited budget. We argue

that large objects are especially informative about the underlying object class and can be labeled at a lower cost than small ones.

- We develop PCIS, a method for training object detectors if only large objects are labeled. This makes it possible to train detectors on extremely sparsely labeled images, e.g. if only 1 out of 200 objects is annotated, and also to approach the performance of a fully labeled dataset with only a subset of the labels.

- We demonstrate that PCIS works better than an oracle that knows the location of every missing annotation and ignores anchors overlapping these regions during training. The oracle can be seen as an upper bound on previous methods described in the literature for dealing with missing annotations [13, 22].

- We show that PCIS outperforms data distillation [14] and unbiased teacher [11], two semi-supervised pseudo-label generation methods if applied to images on which only large objects are labeled.

2 Related Work

In this section, we briefly review representative methods for object detection dealing with different types of incomplete annotations, meaning that not every object has a bounding box label. Those can broadly be divided into three categories, namely weakly-supervised, semi-supervised, and missing annotations. In addition, we cover few-shot object detection, which leverages a small number of labeled images. Finally, we review previous works on rescaling input images in object detection.

In weakly-supervised learning only image-level labels are available but no bounding boxes. That means there is no information about the location, number, size, or aspect ratio of objects. Most of the current methods treat it as a two-step procedure. First, Multiple Instance Learning is used to create pseudo-labels. Second, those are then used to train a fully-supervised detector [27]. Despite recent advances in the field, the quality of the generated pseudo-labels is still much lower than of the actual instance-level labels, resulting in inferior performance [5, 16, 20, 24, 27].

In semi-supervised object detection, some part of the dataset contains instance-level labels, whereas the remaining images are unlabelled. In data distillation [14], pseudo-labels are created by bounding box voting [4] using predictions generated from a baseline model on multiple augmentations of every image. Jeong et al. [6] directly enforce consistency in the model output between an unlabeled image and a flipped version of it using a variant of the Jenson-Shannon divergence. Sohn et al. [19] indirectly enforce consistency by generating pseudo-labels on a weakly augmented image whereas a strongly augmented version of the same image is used in training. Unbiased teacher [11] applies a teacher-student framework in which the teacher generates pseudo-labels that are used to train the student. The teacher is updated using an exponential moving average of the student's weights.

Little research has been conducted on dealing with missing annotations. Thereby, existing works only consider different forms of label sparsity independent of object size. Wu et al. [22] propose to downweight gradients of ROIs based on their overlap with known ground truth objects. This concept was employed by Gao et al. [3] which achieved second place on the public leaderboard in the OpenImages competition 2018. The winning method applied

Pseudo-Label Guided Sampling where regions that are likely to contain a missing annotation were ignored during training [13]. Creating pseudo-labels from box predictions above a confidence threshold that did not strongly overlap known annotations and using them for training slightly degraded performance.

Few-shot object detection aims at adapting a pretrained detector to novel classes for which only a few labeled images are available [9, 17, 23, 25]. Most algorithms follow a two-step procedure. First, a detector is trained on a large, fully-labeled dataset (e.g. 60 out of 80 classes from COCO). Second, this pretrained detector is finetuned on a set of novel classes (e.g. 10 or 30 labeled images for each of the remaining COCO classes).

Overall, annotating only large, easy-to-spot objects has not been covered in previous research. For comparison with our approach, we use three methods from related areas: First, an oracle that confidently ensures that anchors overlapping missing annotations are ignored during training. This can be seen as an upper bound on previous methods dealing with missing annotations [13, 22]. Second, data distillation [14] and unbiased teacher [11], which are pseudo-label methods originally developed for semi-supervised learning. We do not compare our method to few-shot learning algorithms because they address a very different problem. In our case, no large, labeled dataset is available for pretraining. Instead, we have plenty of images, but only a fraction of the largest objects is labeled. This means that a large number of images is unlabeled, whereas the remaining ones are only partially labeled. Standard few-shot learning algorithms do not leverage unlabeled data. If such an algorithm would be applied only on the partially-labeled subset, it would erroneously take up unlabeled objects as negative examples.

Rescaling images for training object detectors has been investigated in various publications [17, 18]. In addition, algorithms for multi-scale inference have been developed [12]. We use downscaling for the specific case of training a detector to find small objects even if only large ones are labeled.

3 Method

3.1 Motivation and Overview

In this section, we propose PCIS, a method specifically designed for training object detectors if only large objects are labeled. The setup is very flexible and can be included in any anchor-based object detector. It contains two main components: First, generating pseudo-labels for missing annotations. These are used for enforcing consistency in the model output across varying input scales. Second, an ignore strategy for anchors, which are likely to overlap missing annotations.

3.2 Pseudo-Labels for Enforcing Consistency in the Model Output

During training, pseudo-labels are generated on the fly by making a prediction on a given training image x using the current model F. The output is post-processed via non-maximum suppression (NMS) and detections above confidence threshold γ, which do not strongly overlap one of the known annotations with class label p^* and coordinates t^* are kept as pseudo-labels. The combination of ground truth annotations and pseudo-labels generated on the image x constitutes the new training targets $Y(p^*, t^*, x)$. A downscaled version of the image x_{scaled} is then used in training (see Fig. 2). This has two effects: First, consistency in the

model output is enforced as the network is trained to detect the pseudo-labeled objects on a smaller version of the same image. This can be seen as a form of consistency regularization. Second, by downsizing known annotations, it is possible to teach the model to detect small objects although they are not annotated in the dataset. While in the beginning of training the network only detects large objects, it gradually learns to find smaller and smaller ones. Pseudo-labels are generated for objects of decreasing size in the process. The combination of pseudo-labels with consistency regularization addresses the aforementioned three main challenges of labeling only large objects. The number of positive examples for each class increases. Assisted by consistency regularization their size gets smaller and smaller. Furthermore, by pseudo-labeling missing annotations, their negative impact on training decreases.

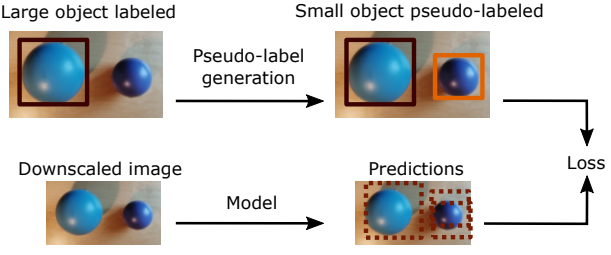

Figure 2: Using pseudo-labels for enforcing consistency in the model output regardless of the input scale. At each point in training, the current model is used to generate pseudo-labels (in orange color) in addition to the available ground truth (in black color). These are then used as training targets for a downscaled and augmented version of the same image. The setup can be seen as a form of consistency regularization.

3.3 Ignore Strategy for Anchors

Previous methods [13] ignore anchors in the loss function based on predictions by a pretrained model, i.e. at locations where the model predicts the presence of an object when there is no corresponding annotation. This has two main downsides. First, if data is very sparsely labeled, the pretrained model performs poorly and hence the regions suggested for ignoring are of little use. Second, difficult background that the pretrained model mistakenly classifies as a missing object gets ignored in training.

To mitigate the negative effect of missing annotations, we take advantage of the usual anchor assignment strategy of anchor-based object detectors. By default, large anchors get matched to large objects and small anchors to small objects. Therefore, each anchor can be seen as an expert for detecting objects of a given size and aspect ratio.

According to our labeling protocol, only large objects are annotated. We call S the size of the anchor that matches the smallest of these labeled objects. As a consequence of the labeling protocol, we know that anchors larger than S do not match missing annotations. Conversely, small objects are not annotated. This means that anchors smaller than S might match a missing annotation. Based on this information, we ignore an individual anchor a in the loss ($I_a = 0$) if its size is smaller than S. Otherwise, we include the respective anchor in the loss ($I_a = 1$). The concept is visualized in Fig. 3. It has the effect that a large fraction of missing annotations is not included in the loss whereas all labeled objects, as well as a large

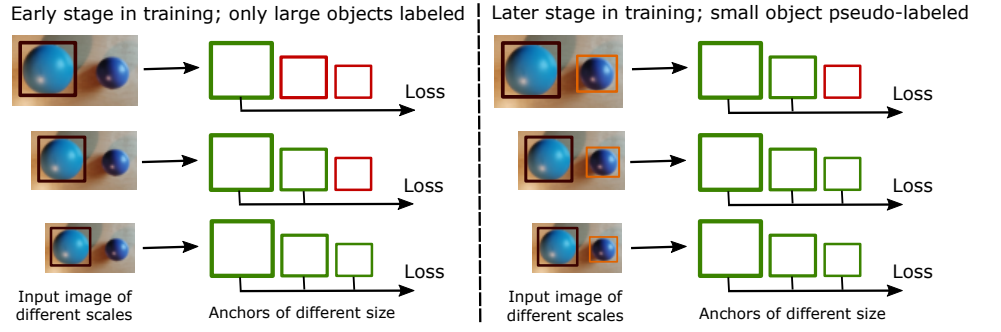

Figure 3: Visualization of the ignore strategy. Anchors are ignored in the loss (in red color) if no anchor of the same size matches a ground truth object (in black color) or pseudo-label (in orange color) and no smaller anchor has a match in the respective image. This reduces the negative effect of missing labels while still keeping a substantial amount of difficult background. By resizing the input image, it is ensured that all anchors are used in training, even if only large objects are labeled. As the model improves during training, more and more pseudo-labels are generated and an increasing share of anchors are included in the loss.

part of difficult background, are kept. In the course of training, an increasing number of pseudo-labels is generated. Therefore, more and more anchor sizes and hence a greater part of the data get included. Overall, this means that in early training, a small but relatively clean subset of the data is used and in later stages, this subset gets gradually larger as pseudo-labels are created.

We develop our formulation using the loss function of the region proposal network head as described in Faster R-CNN [15] for a single image and a single class (see equation 1). The underlying concept can easily be applied to other loss functions used in object detection.

$$\begin{aligned} L(x,p^*,t^*) = &\frac{1}{N_{cls}}\sum_a^A L_{cls}(p_a^*, p_a(x)) \\ &+\frac{\lambda}{N_{reg}}\sum_a^A p_a^* L_{reg}(t_a^*, t_a(x)) \end{aligned} \tag{1}$$

N_{cls} and N_{reg} normalize the classification and regression loss terms L_{cls} and L_{reg} by the number of positive anchors and λ weights them relative to each other. The class and coordinate predictions $p_a(x)$ and $t_a(x)$ are generated using the image x. A denotes the collection of anchors. Equation 2 summarizes the modifications of the loss function made by PCIS.

$$\begin{aligned} L(x,x_{scaled},p^*,t^*) = &\frac{1}{N_{cls}}\sum_a^A I_a L_{cls}(Y_{pa}(p^*,t^*,x), p_a(x_{scaled})) \\ &+\frac{\lambda}{N_{reg}}\sum_a^A I_a Y_{pa}(p^*,t^*,x) L_{reg}(Y_{ta}(p^*,t^*,x), t_a(x_{scaled})) \end{aligned} \tag{2}$$

Fixed training targets p^* and t^* are replaced by a combination of ground truth annotations and pseudo-labels $Y_p(p^*,t^*,x)$ and $Y_t(p^*,t^*,x)$ generated on the input image x. The image is then downscaled (x_{scaled}) to ensure consistency in the model output independent of the image scale. I_a outputs a binary label whether or not to include an anchor in the loss to mitigate the negative effect of missing labels.

4 Experiments

4.1 Datasets

CityPersons is a pedestrian detection dataset. The main evaluation metrics on CityPersons only consider pedestrians above 50 pixels such as the log-average miss rate on the reasonable subset (size ≥ 50 pixels, occlusion ratio < 0.35) which we denote as MR. Therefore, state-of-the-art methods on CityPersons only use pedestrians of size 50 pixels or greater during training and keep the remaining annotations (e.g. small pedestrians, class ignored) as ignore regions [10, 21]. We also follow this convention for computing the upper bound (named 100% largest + Ign.). For the remaining experiments, we discard the ignore regions. This was done because it seemed more plausible that if only limited resources are available for annotating, objects that we want to detect should be labeled first and no effort should be put in annotating ignore regions. This means that in our investigations, 10% of annotations correspond to 10% of pedestrians above 50 pixels and only 4.5% of all annotations.
EuroCityPersons is a diverse pedestrian detection dataset. Like for CityPersons, we only use pedestrians above 50 pixels and discard the remaining annotations. For evaluation, we report log-average miss rate on the reasonable subset.
MS COCO is an object detection dataset. The training split train2017 contains around 115k labeled images and 80 different classes of objects. For evaluation, we report mean average precision (mAP) on the test-dev split.

4.2 Methods for Comparison

We contrast PCIS with four different methods described below. In all experiments, we use a RetinaNet [9] with a ResNet-50 backbone. To ensure a fair comparison, the training settings of PCIS and all competing methods are identical unless specified otherwise.
Supervised baseline: A supervised training protocol is followed and the existence of missing annotations is not taken into account.
Oracle ignore: It is assumed that an oracle is available which confidently detects every missing annotation in the dataset. The corresponding regions of the image are then ignored in the loss function. This can be seen as an upper bound on previous methods for sparsely labeled images [13, 22].
Data distillation [14]: First, the network is trained following the supervised baseline protocol. Pseudo-labels are generated by bounding box voting [4] using multiple scaled and flipped versions of every image in the training set. Finally, the model is retrained using both ground truth annotations and pseudo-labels.
Unbiased teacher [11]: In the burn-in stage, the network is trained following the supervised baseline protocol using the available labels. Afterwards, we follow the mutual teacher-student learning schedule of generating pseudo-labels, retraining the student, and updating the teacher using an exponential moving average of the student's weights.

4.3 Implementation Details

Most of the hyperparameters were chosen according to Lin et al. [9]. During training for PCIS and all the baselines, we used scale jitter in the range 608-1024 for CityPersons and EuroCityPersons as well as 608-800 for MS COCO. For pseudo-label generation in PCIS, we kept the input image size at 1024 for CityPersons and EuroCityPersons and scaled it to

800 for MS COCO. We applied NMS with a threshold λ_{nms} of 0.3 and kept predictions above a confidence γ of 0.5. Pseudo-labels with an IOU of at least 0.6 with a known annotation were removed. Regarding the oracle, we ignored anchors which overlap a missing annotation with an intersection over foreground of 0.3 or more. For pseudo-label generation in data distillation, we scaled the images in the range 842-1323 for Citypersons and EuroCityPersons and well as 400-1200 in MS COCO. The confidence threshold γ was set to 0.5. For unbiased teacher, the confidence threshold γ was set to 0.5 and the EMA rate α to 0.9996.

4.4 Results on CityPersons

Annotations	Supervised	Data dist.	Oracle ign.	Unbiased teacher	PCIS
0.5%	94.0 ± 2.5	-	92.5 ± 2.3	71.2 ± 1.7	$\mathbf{61.2 \pm 1.5}$
1%	91.8 ± 2.4	-	90.2 ± 1.9	46.5 ± 1.5	$\mathbf{41.3 \pm 1.2}$
5%	89.1 ± 1.8	-	74.0 ± 1.4	28.6 ± 1.1	$\mathbf{24.3 \pm 0.9}$
10%	81.4 ± 1.7	72.7 ± 1.6	63.3 ± 1.1	23.0 ± 0.7	$\mathbf{19.9 \pm 0.7}$
25%	61.0 ± 1.3	51.6 ± 1.2	41.5 ± 0.9	20.1 ± 0.7	$\mathbf{17.8 \pm 0.7}$
50%	37.2 ± 1.1	31.5 ± 0.8	21.7 ± 0.5	16.3 ± 0.6	$\mathbf{13.8 \pm 0.5}$
100%	18.4 ± 0.6	-	-	-	-
100% + Ign.	11.7 ± 0.3	-	-	-	-
Training time	13 h	33 h	14 h	23 h	16 h

Table 1: Miss rate on CityPersons for PCIS and all competing methods. The percentage values in the first column denote the share of large objects that are annotated. 'Ign.' indicates that also ignore labels were available. The standard deviation of experiments was computed based on five independent runs. Training time was measured on a single Nvidia Titan RTX when using 50% of the CityPersons annotations.

We evaluated PCIS and the competing methods on CityPersons if only a certain percentage of the largest objects were labeled (see Table 1). The oracle ignore and data distillation improved upon the supervised baseline. Unbiased teacher showed even more promising performance, yet it was outperformed by PCIS for all levels of label sparsity. Using PCIS, a model with a moderate miss rate could be trained if only 0.5% of objects, equivalent to 64 bounding boxes, were annotated. Given just half of the labels, performance close to the lower bound of using 100% of annotations and ignore labels could be achieved. Furthermore, training time was measured on a single Nvidia Titan RTX when using 50% of CityPersons annotations. Data distillation takes the longest among pseudo-label methods because it consists of two trainings as well as computing an offline ensemble for pseudo-label generation. Training unbiased teacher requires more time than PCIS because it uses a moving average of the student's weights for the teacher which only gradually evolves.

4.5 Ablations on CityPersons

Table 2 shows the effect that different components of our method have on performance. As can be seen, not using downscaling is quite detrimental. Without it, generating pseudo-labels has almost no positive effect. If downscaling is applied, the ignore strategy and pseudo-labels independently lead to improvements. Combining them further optimizes results. Table 3 analyses the miss rate of the supervised baseline and PCIS for different object sizes. As can

Method	PL	IS	SC	MR
Supervised	-	-	√	37.2
-	-	-	-	49.1
-	√	-	-	48.5
-	√	√	-	45.7
-	-	√	√	34.2
-	√	-	√	15.7
PCIS	√	√	√	**13.8**

Table 2: Intermediate variants between supervised training and PCIS in case of using the largest 50% of annotations from CityPersons. PL is short for pseudo-label, IS an abbreviation for ignore strategy, SC for (down-)scaling the input image.

		Range of object sizes in evaluation			
Method	MR	50-105	105-166	166-260	>260
50% largest	37.2	63.3	9.4	5.0	7.5
50% largest + PCIS	**13.8**	**16.6**	**6.3**	**2.5**	**6.0**
25% largest	61.0	97.9	46.0	5.4	5.6
25% largest + PCIS	**17.8**	**21.3**	**8.9**	**4.3**	**5.5**
10% largest	81.4	100.0	98.4	40.9	8.6
10% largest + PCIS	**19.9**	**23.9**	**11.6**	**4.1**	**7.3**

Table 3: Miss rate on CityPersons for different object sizes if the largest 50%, 25% or 10% of pedestrians (height $>$ 105 pixels; height $>$ 166 pixels; height $>$ 260 pixels) are labeled. The top row in each section shows results of the supervised baseline, the bottom row of PCIS.

be seen, in the supervised baseline the detector only learns to find large objects but completely fails to detect small ones. Contrary to that, PCIS learns to detect small items quite well, although they are not labeled in the training set.

Overall, the combination of three main components allows PCIS to train powerful object detectors even if only large objects are labeled. Pseudo-labeling unlabeled objects increases the pool of instances to learn from while reducing the number of missing annotations. Enforcing output consistency across scales teaches the model to detect small objects although only large ones are labeled. The anchor scale-dependent ignore strategy mitigates the harmful effect of missing annotations. If just one of these components is missing, performance drops substantially.

4.6 Results on EuroCityPersons

We evaluated PCIS on EuroCityPersons given a varying number of labeled objects. As can be seen in figure 4, PCIS outperformed all other methods. Notably, the gap between PCIS and unbiased teacher was larger than on CityPersons.

4.7 Results on MS COCO

Furthermore, we generated results on MS COCO if only a certain percentage of the largest objects were labeled (see Table 4). PCIS improved upon the supervised baseline, data distil-

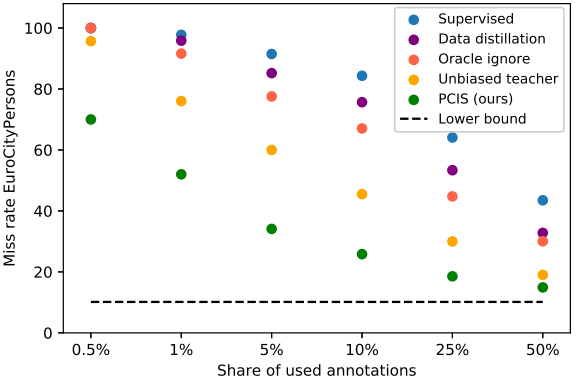

Figure 4: Miss rate on EuroCityPersons by PCIS compared to a supervised baseline, data distillation [14], an oracle ignore, and unbiased teacher [11] if only large objects are labeled.

lation, unbiased teacher, and an oracle ignore. Although the dataset characteristics of COCO are quite different from CityPersons and EuroCityPersons, results for PCIS are robust. The relative performance of the methods was quite similar on all three datasets.

Method	10%	25%	50%	100%
Supervised	4.8	13.4	24.9	35.8
Data distillation	5.7	15.8	26.2	-
Oracle ignore	7.3	17.4	27.2	-
Unbiased teacher	9.3	18.6	27.3	-
PCIS	**10.1**	**21.3**	**27.8**	-

Table 4: Mean average precision (mAP) on MS COCO test-dev by PCIS compared to a supervised baseline, data distillation, an oracle ignore, and unbiased teacher if only large objects are labeled. The percentage at the top row indicates the share of annotated objects.

5 Conclusion

In this work, we propose to rethink labeling strategies for object detection given a limited budget. Specifically, we suggest labeling only large, easy-to-spot objects, which contain more pixels and therefore usually more information about the underlying object class than small ones. To leverage such a dataset, we propose PCIS, a novel combination of pseudo-labels, output consistency across scales, and an anchor scale-dependent ignore strategy. In experiments on CityPersons, EuroCityPersons, and MS COCO, we show that PCIS outperforms an oracle ignoring overlapping anchors as well as competitive pseudo-label generation methods. In summary, PCIS makes it possible to approach the performance of a fully labeled dataset with only a subset of the annotations and to train detectors on extremely sparsely labeled images.

References

[1] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M. Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1844–1861, 2019.

[2] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, and Winston H. Hsu. Dual-awareness attention for few-shot object detection. In *IEEE Transactions on Multimedia TMM*, 2021.

[3] Yuan Gao, Xingyuan Bu, Yang Hu, Hui Shen, Ti Bai, Xubin Li, and Shilei Wen. Solution for large-scale hierarchical object detection datasets with incomplete annotation and data imbalance, 2018. European Conference on Computer Vision ECCV 2018, Open Images workshop.

[4] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1134–1142, 2015.

[5] Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 16797–16807, 2020.

[6] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[7] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7372, June 2021.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. arXiv:1405.0312.

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations ICLR*, 2021.

[12] Mahyar Najibi, Bharat Singh, and Larry S. Davis. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] Yusuke Niitani, Takuya Akiba, Tommi Kerola, Toru Ogawa, Shotaro Sano, and Shuji Suzuki. Sampling techniques for large-scale object detection from sparsely annotated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[16] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G. Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Bharat Singh and Larry S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[19] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection, 2020. arXiv:2005.04757.

[20] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] Wenhao Wang. Adapted center and scale prediction: More stable and more accurate, 2020. arXiv:2002.09053.

[22] Zhe Wu, Navaneeth Bodla, Bharat Singh, Mahyar Najibi, Rama Chellappa, and Larry S. Davis. Soft sampling for robust object detection. In *British Machine Vision Conference BMVC19*, 2019.

[23] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[24] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[25] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning, 2021. arXiv:2103.11731.

[26] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection, 2017. arXiv:1702.05693.

[27] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.