ClueNet : A Deep Framework for Occluded Pedestrian Pose Estimation

Perla Sai Raj Kishore*[1]
sairajkishore13@gmail.com
Sudip Das*[2]
d.sudip47@gmail.com
Partha Sarathi Mukherjee[2]
pathosarothimukherjee@gmail.com
Ujjwal Bhattacharya[2]
ujjwal@isical.ac.in

[1] Institute of Engineering & Management, Kolkata, India
[2] Indian Statistical Institute, Computer Vision & Pattern Recognition Unit, Kolkata, India
* indicates equal contributions

Abstract

Pose estimation of a pedestrian helps to gather information about the current activity or the instant behaviour of the subject. Such information is useful for autonomous vehicles, augmented reality, video surveillance, etc. Although a large volume of pedestrian detection studies are available in the literature, detection of the same in situations of significant occlusions still remains a challenging task. In this work, we take a step further to propose a novel deep learning framework, called ClueNet, to detect as well as estimate the *entire pose* of occluded pedestrians in an unsupervised manner. ClueNet is a two stage framework where the first stage generates visual clues for the second stage to accurately estimate the pose of occluded pedestrians. The first stage employs a multi-task network to segment the visible parts and predict a bounding box enclosing the visible and occluded regions for each pedestrian. The second stage uses these predictions from the first stage for pose estimation. Here we propose a novel strategy, called *Mask and Predict*, to train our ClueNet to estimate the pose even for occluded regions. Additionally, we make use of various other training strategies to further improve our results. The proposed work is first of its kind and the experimental results on CityPersons and MS COCO datasets show the superior performance of our approach over existing methods.

1 Introduction

Since last several years pedestrian detection has been one of the interesting and challenging tasks in the Computer Vision community. Although its state-of-the-art methods have now achieved significant accuracies, pose estimation of pedestrians, particularly for partially occluded pedestrians still remains a challenging problem. Precise estimation of pose not only helps to identify a pedestrian more accurately but also to gather information about the present activity of the subject or its instant behaviour. Automatic understanding of limb articulation or posture of pedestrians is useful for autonomous vehicles, robotics, augmented reality, video surveillance, etc. For the inherent complexity of the problem, mainly due to

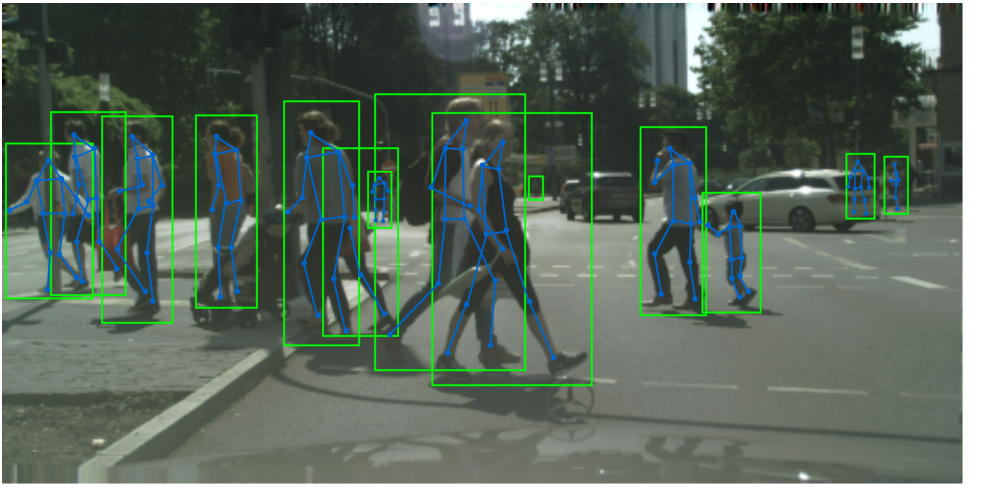

Figure 1: Pedestrian detection and pose estimation results of ClueNet on an image sample of CityPersons benchmark dataset.

the high variation in scales of the figures appearing in the vicinity or different degrees of occlusion caused by various obstacles in the scene, it still remains a challenging research problem. Although during the last decade major progress has been achieved on pedestrian detection [4, 14, 54, 56], their performance in cases of occlusion drops significantly [53]. Occlusion in pedestrian figures is a common phenomenon but enough studies for efficient handling of similar cases cannot be found in the literature. A majority of related studies [34, 43, 59] considered some ensemble methods and consequently their executions are time intensive. On the other hand, to the best of our knowledge, no such study has been conducted on pose estimation of occluded pedestrians.

Early attempts of pose estimation were based on robust image feature computation and use of some sophisticated classification framework [48]. However, all of the recent studies [46] use convolutional networks and do not indulge in explicit feature computation although a majority of these approaches fails to perform efficiently in cases of occlusions. Some improvements in cases of occlusions have been reported in a few studies [11, 23, 30]. In this work, we have introduced ClueNet, a novel two stage deep learning based framework to address this problem and the experimental results on benchmark datasets show significant improvement over state-of-the-art models.

The proposed ClueNet framework consists of two stages. The first stage aims to segment the visible parts of the pedestrian instances and localize the pedestrians by bounding boxes, enclosing both it's visible and occluded regions, through a two stream multi-task scale invariant network. The segmentation and detection streams share a common feature encoder network that uses ResNet-50 [19] and modified Inception blocks [41] with vertically shaped convolution kernels to segment and detect pedestrians of different scales in an efficient manner. We further use channel-wise attention to enhance important features, while at the same time attenuate noisy and unwanted background information. The multi-task setting allows these correlated tasks (segmentation and detection) to learn from each other and hence perform better than they would if used individually. The segmentation and detection results, which act as visual clues, are then passed to the second stage of the ClueNet for pose estimation.

The second stage of our framework, which is also our primary contribution, is a new pose estimation model that takes input from the predictions of the first stage and aims to estimate the *entire pose* of both completely visible as well as occluded pedestrians. But again, estimating the entire pose for occluded pedestrians is a challenging task due to few reasons. First, none of the existing pedestrian detection datasets come with annotations for the pose of pedestrians. Secondly, even if we consider using any 2D pose estimation dataset alongside, none of these datasets come with annotations for pose in the occluded regions. To tackle this, we propose a novel training procedure which makes use of Domain Adaptation [45] and the proposed *Mask and Predict* strategy to train the model to estimate the *entire pose* of a pedestrian (irrespective of complete visibility) in an unsupervised way. To the best of our knowledge this is the first approach to explore the task of estimating the entire pose of occluded pedestrians, or human instances in general.

2 Related Work

The proposed solution of occluded pedestrian pose estimation task consists of three subtasks, *viz.* detection, semantic segmentation and pose estimation of pedestrian figures in a scene image. Pedestrian detection has been studied extensively [1, 3, 8, 25, 60] during the past decade. However, occluded pedestrian detection remains a challenge even in the present day. Existing studies of Occluded Pedestrian Detection [35, 36, 42, 59] aimed to learn the various occlusion patterns available in the training set. Zhou *et al*. [59] proposed a part detector method to learn the occlusion patterns using a multi-label learning approach. Such methods fail to generalize effectively in real world occlusion scenarios. Although performance of vanilla Faster R-CNN [38] gets suffered on smaller pedestrian figures, the SSD model proposed in [29] significantly outperformed the performance of Faster R-CNN in terms of both accuracy and speed. Later, He *et al*. [31] took the help of Region Proposal Network [38] and Boosted Forests [7] to improve the detection performance of smaller pedestrian figures. More recently, Zhang *et al*. [53] used a regression network with guided attention improving pedestrian detection performance for different types of occlusion. Segmentation can either be based on individual instances (Instance Segmentation) or on a broader category or type of the objects (Semantic Segmentation). Recent studies of segmentation used CNN based frameworks and some of these include [9, 10, 16, 28, 50] and [62].

State-of-the-art keypoint estimation performance could be obtained by using convolutional neural networks (CNN). Such an architecture proposed by Alejandro *et al*. [31] grouped the features of different scales without losing spatially correlated information. Performance of a few other real-time pose estimation frameworks [6, 7] is also satisfactory. The model proposed by Kocabas *et al*. [24] handles person detection, segmentation and its pose estimation. Ke *et al*. [40] studied the human pose estimation based on learning high resolution representations. Pengfei *et al*. [52] proposed a semantics-guided neural network (SGN) for action recognition. Rohit *et al*. [17] used a two stage architecture to estimate pose and track the same in a complex scene. Recently, Ye and Kim [49] have studied a similar problem in 3D hand pose estimation task. On the other hand, the Occlusion-Net [37] proposed by Reddy *et al*. used graph networks to predict 2D and 3D keypoint locations of the occluded parts of non-human objects while Zhu *et al*. [61] used Occlusion-adaptive Deep Networks to solve facial landmark detection problems. Hou *et al*. [20] proposed a Spatio-Temporal Completion network to tackle occlusion in Person Re-Identification by generating the contents of occluded parts using spatial structure and temporal information of pedestrian sequences.

3 Proposed Methodology

The proposed framework is divided into two stages. The first stage aims to detect and segment pedestrians from the input image whereas the second stage estimates the entire pose of the detected pedestrians. The following sections describe these two stages in detail.

3.1 Stage 1: Pedestrian Detection and Segmentation

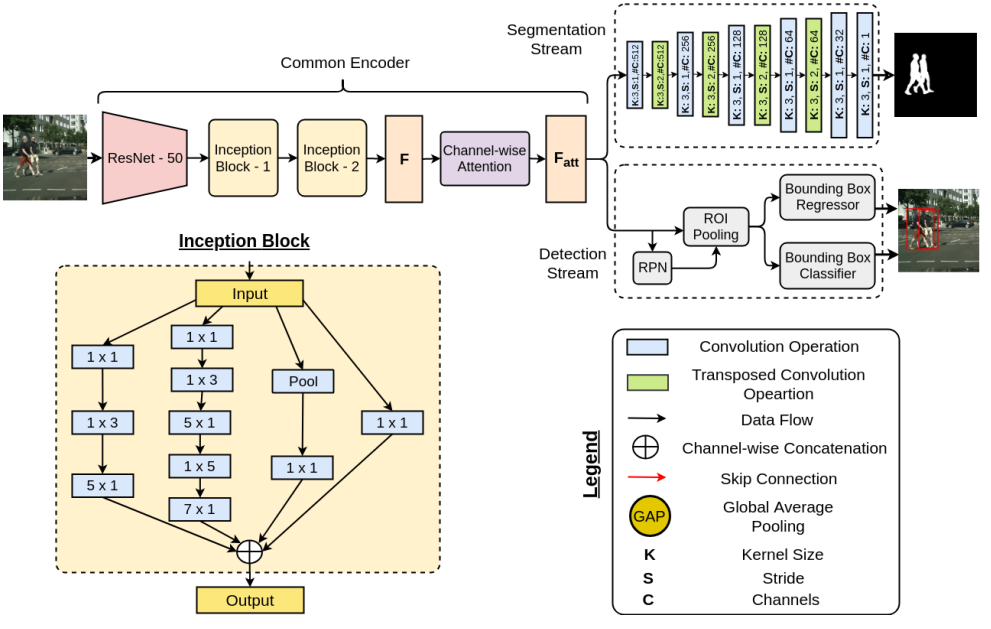

Figure 2: Architecture of Stage 1 of the proposed ClueNet framework.

The first stage of the proposed framework aims to segment and detect all the pedestrians in the scene, both completely visible and occluded, from the input image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$. We design a novel two stream multi-task scale invariant network which can segment and detect pedestrians of different scales efficiently. The model consists of a common feature encoder followed by two task specific convolutional streams for segmentation and detection. Multi-task learning allows the segmentation and detection streams to learn collaboratively and capitalize on one another's resources and skills. The architecture of the proposed model is shown in Figure 2. The common feature encoder consists of convolution layers of ResNet-50 up till `conv5_9` followed by two Inception v3 [41] blocks modified to handle pedestrians of different scales. Different branches of the inception block have different receptive fields which help the task networks to segment and detect pedestrians of various scales in an efficient manner. Inspired from the usual shape of pedestrians, we use vertically shaped filters in the inception blocks instead of traditionally used square ones. This allows us to extract more information from the vertical direction than horizontal at one go, which is more suitable for a task like pedestrian detection or segmentation. Such an approach reduces memory footprint and also eliminates the extra information from horizontal direction, making the convolution operations more efficient and effective.

The final part of the common encoder is a channel-wise attention module that attends to various channels of the feature map to enhance the important information, while attenuating noise and background information before passing it to task specific layers of the model. We first pass the feature map $\mathcal{F}$ from the last inception block through a 1×1 convolution layer to generate $\mathcal{F}_{1\times1}$ of same size and number of channels. $\mathcal{F}_{1\times1}$ is then passed through sigmoid activation function and multiplied with $\mathcal{F}$ element-wise to generate the attended feature map $\mathcal{F}_{att}$. Mathematically,

$$\mathcal{F}_{att} = \mathcal{F} \otimes \sigma(Conv_{1\times1}(\mathcal{F})) \tag{1}$$

where $Conv_{1\times1}(\cdot)$ represents a 1×1 convolution operation, $\sigma(\cdot)$ is the sigmoid activation function and $\otimes$ represents element-wise multiplication.

The segmentation specific network is a decoder with stacked convolution and transposed convolution layers as shown in Figure 2. Here we aim to classify every pixel of $\mathcal{I}$ to one of two categories: *Pedestrian* or *Background*. The network takes $\mathcal{F}_{att}$ as input and generates a single channel binary mask $\widehat{\mathcal{M}} \in \mathbb{R}^{H\times W\times1}$, passed through sigmoid activation, as output. The segmentation network is trained using Binary Cross Entropy (BCE) Loss, $\mathcal{L}_{seg}$, with respect to the ground truth mask $\mathcal{M}$ where the white pixels represent *Pedestrians* and black pixels represent *Background*.

The detection specific network is inspired from Faster-RCNN [38] and consists of a Regional Proposal Network (RPN) followed by bounding box regression and classification networks. The RPN generates object proposals which are then more accurately localized and classified as *Pedestrian* or *Backdround* by the following networks. The associated loss function $\mathcal{L}_{det}$, is same as used in the original paper [38], smooth $\mathcal{L}_1$ loss for regressor and BCE loss for classifier. The overall loss function for Stage 1, $\mathcal{L}_{Stage1}$, is given as the weighted combination of the two task losses as,

$$\mathcal{L}_{Stage1} = \mathcal{L}_{seg} + \lambda\,\mathcal{L}_{det} \tag{2}$$

3.2 Stage 2: Pose Estimation

The second stage of the proposed framework is a pose estimation network that is trained to estimate the entire pose of visible as well as occluded pedestrians in an unsupervised manner. As mentioned earlier, we do not get annotations for human pose in the occluded regions from any of the existing pedestrian detection or pose estimation datasets. To solve this, we design a novel solution that makes use of the following two strategies:

Domain Adaptation: Since pose annotations for pedestrian detection datasets are not available, we aim to learn estimating poses from an additional pose estimation dataset and use this knowledge to estimate the poses of pedestrians. However, due to shift in the distribution of data, the model trained to estimate poses on pose estimation dataset does not work well for the unseen pedestrian detection dataset. To tackle this, we use an unsupervised adversarial domain adaptation approach [44] that minimizes the domain shift between datasets to generate better results on pedestrian detection dataset.

Mask and Predict: This strategy attempts to solve the problem of estimating the entire pose of occluded pedestrians. By this strategy, we first randomly mask certain regions of the person whose pose is to be estimated by black patches of different shapes and sizes, and then ask the network to predict the pose for the visible as well as the masked regions. Masking here is analogous to occlusion and hence the model learns to estimate plausible poses for even occluded pedestrians. Random masking allows the network to efficiently learn even

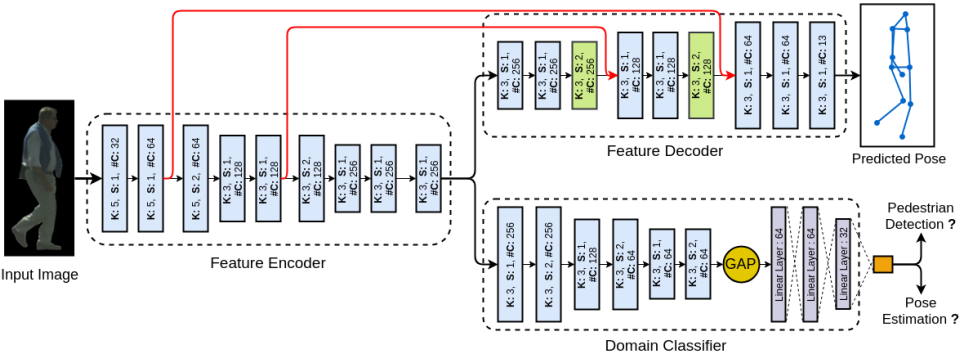

Figure 3: Architecture of Stage 2 of the proposed ClueNet framework. The legend for this image is provided in Figure 2.

unseen occlusion patterns during training, making the model more robust and thus generalize well during testing.

The proposed second stage uses three networks, a feature encoder $\mathcal{E}(\cdot)$, a decoder $\mathcal{D}(\cdot)$ and a domain classifier $\mathcal{C}(\cdot)$. The architecture of the proposed model is shown in Figure 3. $\mathcal{E}(\cdot)$ takes individual human instances as input and generates dense features as output, $\mathcal{D}(\cdot)$ takes these dense features and estimates the pose by generating a heat map for every keypoint we want to estimate and $\mathcal{C}(\cdot)$ along with $\mathcal{E}(\cdot)$ is trained to reduce the domain shift between the pose estimation and pedestrian detection datasets. To handle the problem of domain shift between datasets we employ an adversarial approach, where $\mathcal{E}(\cdot)$ is analogous to the Generator of a Generative Adversarial Network (GAN) [13] and $\mathcal{C}(\cdot)$ is analogous to the Discriminator. $\mathcal{E}(\cdot)$ here aims to generate dataset invariant features for input human instances from the two datasets while $\mathcal{C}(\cdot)$ aims to classify these features as to which dataset they belonged to. The $\langle \mathcal{E}(\cdot), \mathcal{D}(\cdot) \rangle$ and $\langle \mathcal{E}(\cdot), \mathcal{C}(\cdot) \rangle$ pairs are trained alternatively to estimate the pose of human instances from the pose estimation dataset and reduce the domain shift among the two datasets respectively. Upon convergence, $\mathcal{E}(\cdot)$ learns to generate dataset invariant features for the input human instances and $\mathcal{D}(\cdot)$ learns to estimate poses from these features, as a result, we can now estimate the pose of human instances from the pedestrian detection dataset with higher accuracy.

Formally, let $\mathcal{X}_{PE}$ be the input images and $\mathcal{Y}_{PE}$ be the corresponding keypoint labels of human instances from the pose estimation dataset (source domain) with distribution $\mathcal{P}_{PE}(x,y)$. Similarly, let $\mathcal{X}_{PD}$ be the input images of human instances, with no corresponding keypoint labels, from the pedestrian detection dataset (target domain) with distribution $\mathcal{P}_{PD}(x,y)$. We then aim to minimize the following objective functions for Domain Adaptation,

$$\begin{aligned} \min_{\mathcal{C}} \ \mathcal{L}_{adv_{\mathcal{C}}}(\mathcal{X}_{PE}, \mathcal{X}_{PD}, \mathcal{E}) &= -\mathbb{E}_{x_{PE} \sim \mathcal{X}_{PE}}[\log \mathcal{C}(\mathcal{E}(x_{PE}))] - \mathbb{E}_{x_{PD} \sim \mathcal{X}_{PD}}[\log(1 - \mathcal{C}(\mathcal{E}(x_{PD})))] \\ \min_{\mathcal{E}} \ \mathcal{L}_{adv_{\mathcal{E}}}(\mathcal{X}_{PE}, \mathcal{X}_{PD}, \mathcal{C}) &= -\mathbb{E}_{x_{PD} \sim \mathcal{X}_{PD}}[\log \mathcal{C}(\mathcal{E}(x_{PD}))] \end{aligned} \tag{3}$$

Similar to [24], we predict a heatmap for each keypoint we want to estimate which represent the keypoint locations as Gaussian peaks for pose estimation. Correspondingly, the ground-truths are also modified with Gaussian peaks at keypoint locations. We then simply use $\mathcal{L}_2$ loss to calculate the error in prediction,

$$\min_{\mathcal{E}, \mathcal{D}} \ \mathcal{L}_{Pose}(\mathcal{X}_{PE}, \mathcal{Y}_{PE}) = -\mathbb{E}_{(x_{PE}, y_{PE}) \sim (\mathcal{X}_{PE}, \mathcal{Y}_{PE})} |\mathcal{D}(\mathcal{E}(x_{PE})) - y_{PE}|_2^2 \tag{4}$$

We use cropped and segmented human instances as input to the second stage network. Cropped instances significantly minimize the false-positive rate and reduce the amount of memory and time required to process each image, we use bounding box predictions from the first stage for this purpose. During training, we follow the *Mask and Predict* strategy by which we mask the input by black patches of different sizes and shapes to artificially occlude the human instances. But at inference, occlusion could be due to a wide a variety of objects like cars, vegetation, poles etc. So, the distribution of input images while training and inference changes. To tackle this, we segment the input images to subtract the background and just keep the human figures. By segmenting the image, we not only subtract the background but also the inanimate objects that occlude the human instances, and hence making the model invariant to the various objects that occlude them. This also helps to estimate the pose easily and more accurately. By masking such a segmented image, we do not change the distribution of the images at the time of training and inference. Segmentation is efficiently done by simply multiplying the original input image with it's segmentation mask, $\widehat{\mathcal{M}}$, generated in the first stage. Furthermore, as we want to estimate the entire pose of the occluded human instances, we only use completely visible individual human instances from the pose estimation dataset for training. These human instances are then segmented, artificially occluded and fed to the network for predicting the entire pose of the person. Since we choose completely visible human instances for this purpose, we have ground-truth for the entire pose with respect to which the error in prediction is calculated and backpropagated through the network.

4 Experimentation Details

4.1 Datasets and Evaluation Metrics

To train our complete system we use two different datasets, CityPersons [53] and Microsoft COCO [27]. The main purposes of the CityPersons dataset is for pedestrian detection and semantic segmentation. It consists of 2,975 training, 500 validation and 1,575 testing images. Microsoft COCO contains the annotation of key points for pose estimation, bounding boxes for objects detection, segmentation mask for scene image understanding and image captioning. We use key point annotations and segmentation masks from this dataset for training the second stage of our framework. COCO dataset has annotations for 17 key points but we consider only 13 key points for our work. They are left and right elbow, shoulder, wrist, hip, knee and ankle along with the head.

We use log average Miss Rate (MR) to calculate the error in pedestrian detection. For pose estimation we experiment with different occlusion scenarios which are Reasonable (R) with $[.65, inf]$ visibility, Heavy occlusion (HO) with $[.20, .65]$ visibility and Reasonable + Heavy occlusion (R + HO) with $[.20, inf]$ visibility. We use IoU (Intersection over Union) to quantify the segmentation results. Average Precision (AP) is used to show the performance for pose estimation in the second stage of the framework.

4.2 Training Details

In the first stage, we use ResNet-50 backbone followed by two modified Inception v3 blocks as a feature extractor. The weights for ResNet-50 layers are initialized by pre-training it on ImageNet [13] and that of other layers are initialized from a truncated normal distribution

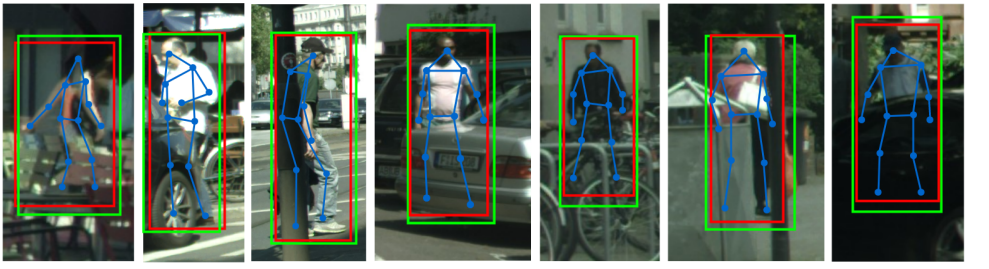

Figure 4: Qualitative results of Detection and Pose Estimation on CityPersons dataset. The ground truth annotations are shown in red, the detection results are shown in green and the predicted pose is shown in blue.

with mean 0 and standard deviation 0.1. Instead of training all layers at once, we make use gradual unfreezing [21] to retain previous knowledge and avoid catastrophic forgetting in the pre-trained ResNet layers. We first train the network by freezing the ResNet layers for $10k$ iterations after which the ResNet layers were gradually unfreezed every few iterations from the last layers and trained jointly with the previously unfreezed layers until we finetune all layers till convergence. The hyper parameter λ is set to 1 and the network is trained using Momentum Optimizer with an initial learning rate of 0.01 and momentum of 0.9. The learning rate is reduced by a factor of 10 after each subsequent $15k$ iterations.

In the second stage, we randomly mask the input images by rectangular patches of various sizes and aspect ratios to artificially occlude the human instances while training. We use Curriculum Learning [5] for this purpose, which aims to gradually increment the complexity of the data fed to the input neural network. In this context, we feed the second stage network with masked input images with masking % gradually increasing from 0% to 70% over training. Since we make use of domain adaptation to estimate the pose of the pedestrians, it is important to carefully choose the source dataset from which we aim to transfer the pose estimation knowledge from. We must ensure that the distribution of the pose (*i.e.* the output) of pedestrians and that of the source dataset should be similar to each other, failing which can lead to inappropriate and erroneous pose predictions on the pedestrian detection dataset. We use MS COCO as our source dataset, which has human instances with poses similar to that of pedestrians, as a result of which the trained model generalizes well on the CityPersons dataset. Data augmentation is employed in both the stages by randomly scaling images between 0.8 to 1.2 and horizontally flipping images with a probability of 0.3. The optimizer and learning rate for second stage are similar to that used in the first stage. All experiments were performed on two Nvidia P6 GPUs.

4.3 Evaluation results

In this subsection, we report our experimental results on CityPersons and MS COCO datasets with a comparison to various other state-of-the-art methods. In the first stage, we attain state-of-the-art results on both Segmentation and Detection tasks. Our model achieves **89.3%** accuracy in terms of IoU for person segmentation and an overall Miss-Rate of **30.84%** on CityPersons dataset. We also achieve, state-of-the-art detection results on Heavily Occluded (HO) pedestrians with a Miss-Rate of **47.68%**. A more detailed outcome of our experiments for Segmentation and Detection are shown in Table 1 and 2 respectively. For the second stage, we evaluate the pose estimation model on MS COCO dataset and the results

Figure 5: Qualitative results of segmentation on Citypersons dataset from the first stage of the network.

are shown in Table 4. The AP here is calculated only for the annotated keypoints of the human instances. Since we do not have any ground-truth annotations for the pose of pedestrians in the CityPersons dataset, we calculate the percentage of the predicted keypoints inside the ground truth segmentation mask for full pedestrian instances as an evaluation metric. We observe an accuracy of **90.18%** on this metric. Additionally, to show the significance of our *Mask and Predict* strategy we perform experiments with occlusion masks of different sizes and the corresponding results on full human instances from MS COCO dataset are shown in Table 3. Some qualitative results of detection and pose estimation on CityPersons dataset are shown in Figure 1 and 4, and that of segmentation are shown in Figure 5. The quantitative and qualitative results show significant improvement in occluded pedestrian detection and pose estimation in different scenarios.

Table 1: Different Benchmark segmentation results on CityPersons dataset.

Model	IoU
DeepLab [9]	79.8
Piecewise [26]	81.5
PSPNet [58]	86.5
DenseASPP [47]	86.2
DANet [15]	87.3
Ours	**89.3**

Table 2: MR for different SOTA models on CityPersons dataset.

Model	R	HO	R+HO
Faster RCNN [38]	15.52	64.83	41.45
Shanshan *et al.* [53]	15.96	56.66	38.23
Junhyug *et al.* [33]	16.77	48.52	31.72
Tao *et al.* [39]	14.4	52.0	34.24
OR-CNN [57]	**11.0**	51.0	36.11
Ours	11.87	**47.68**	**30.84**

Table 3: Results on MS COCO dataset with different occlusion percentages.

Occlusion Percentage	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**
Average Precision	88.06	83.93	79.8	73.4	64.0	58.8

Table 4: AP for state-of-the-art models on MS COCO dataset.

Model	AP	AP_{50}	AP_{75}	AP_M	AP_L
Newell *et al.* [32]	65.5	86.8	72.3	60.6	72.6
CMU-Pose[6]	61.8	84.9	67.5	57.1	68.2
MultiPoseNet [24]	69.6	86.3	76.6	65.0	76.3
Megvii [12]	73.0	**91.7**	**80.9**	69.5	78.1
CFN [22]	72.6	86.7	69.7	**78.3**	79.0
Ours	**73.9**	89.6	78.2	70.9	**79.1**

5 Conclusion

In this work we proposed a novel framework to accurately detect and estimate the entire pose of completely visible as well as occluded pedestrians. We used a two stage framework for this purpose, the first stage employs a multi-task network to detect and segment pedestrians from the input image and the second stage estimates the entire pose of the detected pedestrians in an unsupervised manner. Various training strategies such as Gradual Unfreezing, Domain Adaptation, *Mask and Predict*, and Curriculum Learning are used to further improve our results which in case of Detection, Segmentation, and Pose Estimation show the superior performance of our approach over existing techniques. More specifically, in this work we used semantic segmentation to subtract the surroundings and leave alone the human figures as input to the second stage. This is to make the model invariant to various objects that occlude the pedestrians, but semantic segmentation fails to handle intra-class occlusions (*i.e.* person-person occlusions). As a result, the model misses out on predicting the *entire pose* in such cases and predicts keypoints only for the visible parts of the occluded pedestrians. We believe the results in such cases can be improved with instance segmentation. However, the proposed work is first of its kind with several real world applications, and can be adopted for other similar occluded object detection and pose estimation tasks easily.

References

[1] Anelia Angelova, Alex Krizhevsky, Vincent Vanhoucke, Abhijit Ogale, and Dave Ferguson. Real-time pedestrian detection with deep network cascades. 2015.

[2] Ron Appel, Thomas Fuchs, Piotr Dollár, and Pietro Perona. Quickly boosting decision trees–pruning underachieving features early. In *International Conf. on machine learning*, pages 594–602, 2013.

[3] Rodrigo Benenson, Markus Mathias, Radu Timofte, and Luc Van Gool. Pedestrian detection at 100 frames per second. In *2012 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2903–2910. IEEE, 2012.

[4] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conf. on Computer Vision*, pages 613–627. Springer, 2014.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proc. of the 26th annual international Conf. on machine learning*, pages 41–48. ACM, 2009.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[8] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 734–750, 2018.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 801–818, 2018.

[11] Xianjie Chen and Alan L. Yuille. Parsing occluded people by flexible compositions. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conf. on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2012.

[15] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018.

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 350–359, 2018.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[21] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[22] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *Proc. of the IEEE International Conf. on Computer Vision*, pages 3028–3037, 2017.

[23] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *14th European Conf. on Computer Vision (ECCV) 2016*, pages 34–50, 2016.

[24] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 417–433, 2018.

[25] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 732–747, 2018.

[26] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conf. on computer vision*, pages 740–755. Springer, 2014.

[28] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. *arXiv preprint arXiv:1901.02985*, 2019.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conf. on computer vision*, pages 21–37. Springer, 2016.

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *14th European Conf. on Computer Vision (ECCV)*, pages 483–499, 2016.

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conf. on Computer Vision*, pages 483–499. Springer, 2016.

[32] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017.

[33] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 966–974, 2018.

[34] Wanli Ouyang and Xiaogang Wang. Joint deep learning for pedestrian detection. In *Proc. of IEEE International Conf. on Computer Vision*, ICCV '13, pages 2056–2063. IEEE Computer Society, 2013.

[35] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1874–1887, 2018.

[36] Bojan Pepikj, Michael Stark, Peter Gehler, and Bernt Schiele. Occlusion patterns for object class detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3286–3293, 2013.

[37] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[38] 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*, pages 91–99, 2015.

[39] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 536–551, 2018.

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pages 2818–2826, 2016.

[42] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele. Detection and tracking of occluded people. *International Journal of Computer Vision*, 110(1):58–69, 2014.

[43] Y Tian, P Luo, X Wang, and X Tang. Deep learning strong parts for pedestrian detection. *Proc. IEEE Int. Conf. On Computer Vision*, pages 1904–1912, 01 2015.

[44] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[45] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[46] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732. IEEE Computer Society, 2016.

[47] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018.

[48] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2878–2890.

[49] Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 801–817, 2018.

[50] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[51] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conf. on computer vision*, pages 443–457. Springer, 2016.

[52] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. *arXiv preprint arXiv:1904.01189*, 2019.

[53] S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in CNNs. In *Proceedings of CVPR*, pages 6995–7003. IEEE, 2018.

[54] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1259–1267, 2016.

[55] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.

[56] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:973–986, 2018.

[57] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 637–653, 2018.

[58] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pages 2881–2890, 2017.

[59] Chunluan Zhou and Junsong Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *The IEEE International Conf. on Computer Vision (ICCV)*, Oct 2017.

[60] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *Proc. of the European Conf. on Computer Vision (ECCV)*, pages 135–151, 2018.

[61] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3486–3496, 2019.

[62] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proc. of the IEEE Conf. on computer vision and pattern recognition*, pages 8697–8710, 2018.