PedHunter: Occlusion Robust Pedestrian Detector in Crowded Scenes

Cheng Chi[1,3*], **Shifeng Zhang**[2,3*], **Junliang Xing**[2,3], **Zhen Lei**[2,3], **Stan Z. Li**[2,3], **Xudong Zou**[1,3]
[1] Institute of Electronics, Chinese Academy of Sciences, Beijing, China
[2]CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China

Abstract

Pedestrian detection in crowded scenes is a challenging problem, because occlusion happens frequently among different pedestrians. In this paper, we propose an effective and efficient detection network to hunt pedestrians in crowd scenes. The proposed method, namely PedHunter, introduces strong occlusion handling ability to existing region-based detection networks without bringing extra computations in the inference stage. Specifically, we design a mask-guided module to leverage the head information to enhance the feature representation learning of the backbone network. Moreover, we develop a strict classification criterion by improving the quality of positive samples during training to eliminate common false positives of pedestrian detection in crowded scenes. Besides, we present an occlusion-simulated data augmentation to enrich the pattern and quantity of occlusion samples to improve the occlusion robustness. As a consequent, we achieve state-of-the-art results on three pedestrian detection datasets including CityPersons, Caltech-USA and CrowdHuman. To facilitate further studies on the occluded pedestrian detection in surveillance scenes, we release a new pedestrian dataset, called SUR-PED, with a total of over $162k$ high-quality manually labeled instances in $10k$ images. The proposed dataset, source codes and trained models will be released.

Introduction

Pedestrian detection is an important research topic in computer vision with various applications, such as autonomous driving, video surveillance and robotics. The goal of pedestrian detection is to predict a bounding box for each pedestrian instance in an image. Although great progress has been made in the realm of pedestrian detection (Tian et al. 2015b; Zhang, Benenson, and Schiele 2015; Costea and Nedevschi 2016; Du et al. 2017; Lin et al. 2018; Noh et al. 2018), occlusion still remains as one of the most challenging issues. Some efforts have been made particularly to handle the occlusion problem, but a notable amount of extra computations is introduced and there is still much room to improve the detection performance.

In this work, we propose a high-performance pedestrian detector namely PedHunter to improve occlusion robustness in crowded scenes, which mitigates the impact of occlusion without sacrificing inference speed. Firstly, inspired by human visual system that often resorts to the head information to locate each pedestrian in crowded scenes, we propose a mask-guided module to predict head masks that helps to enhance the representation learning of pedestrian features during training. The mask-guided module does not participate in inference, so it improves the accuracy without computational overhead. Secondly, we observe that error detections between occluded pedestrians are the common false positive in crowd scenes, which results from the low-quality positive samples during training. To improve the quality of positive samples, we develop a strict classification criterion by increasing the IoU threshold of positive samples and adding jittered ground truths, which reduces aforementioned false positives. Thirdly, we introduce an occlusion-simulated data augmentation technique that generates random occlusion on ground truths during training. It significantly enriches the pattern and quantity of occlusion, making the proposed model more robust to occlusion.

In addition, current pedestrian detection benchmarks usually focus on autonomous driving scenes. It is necessary to build a new benchmark in the field of surveillance, which is another important application domain for pedestrian detection. To this end, we introduce a new SUR-PED dataset for occluded pedestrian detection in surveillance scenes. It has over $162k$ pedestrian instances manually labeled in $10k$ surveillance camera images. Based on this proposed dataset and exiting datasets including CrowdHuman (Shao et al. 2018), extented CityPersons (Zhang, Benenson, and Schiele 2017) and extented Caltech-USA (Dollár et al. 2009)[1], several experiments are conducted to demonstrate the superiority of the proposed method, especially for the crowded scenes. Notably, the proposed PedHunter detector achieves state-of-the-art results without adding any extra overhead, *i.e.*, 8.32% MR^{-2} on CityPersons, 2.31% MR^{-2} on Caltech-USA and 39.5% MR^{-2} on CrowdHuman.

To summarize, the main contributions of this work are in five-fold as follows: 1) Proposing a mask-guided module to enhance the discrimination ability of the occluded features;

*These authors contributed equally to this work.

[1]In our previous work, we additionally label the corresponding head bounding box for each annotated pedestrian instance as the extended version of CityPersons and Caltech-USA.

2) Designing a strict classification criterion to provide higher quality positives for R-CNN to reduce false positives; 3) Developing an occlusion-simulated data augmentation to enrich occlusion diversity for stronger robustness; 4) Providing a new surveillance pedestrian dataset to facilitate further studies on occluded pedestrian detection; 5) Achieving state-of-the-art results on common pedestrian detection datasets without adding additional overhead.

Related Work

Pedestrian detection is dominated by CNN-based methods (Hosang et al. 2015; Yang et al. 2015) in recent years. Sermanet *et al.* (Sermanet et al. 2013) use the convolutional sparse coding to pre-train CNN for pedestrian detection. Cai *et al.* (Cai, Saberian, and Vasconcelos 2015) persent a complexity-aware cascaded detector for an optimal trade-off between accuracy and speed. Angelova *et al.* (Angelova et al. 2015) detect pedestrian by combining the ideas of fast cascade and a deep network. Zhang *et al.* (Zhang et al. 2016a) present an effective pipeline for pedestrian detection via using RPN followed by boosted forests. Mao *et al.* (Mao et al. 2017) introduce a novel network architecture to jointly learn pedestrian detection with the given extra features. Li *et al.* (Li et al. 2018) use multiple built-in sub-networks to adaptively detect pedestrians across scales. Brazil *et al.* (Brazil, Yin, and Liu 2017) exploit weakly annotated boxes via a segmentation infusion network to achieve considerable performance gains.

Although significant progresses have been made from CNN-based pedestrian methods, it remains a very challenging problem to detect occluded pedestrian in crowd scenes. Several methods (Tian et al. 2015a) describe the pedestrian using part-based model to handle occlusion, which learn a series of part detectors and design some mechanisms to fuse the part detection results to localize partially occluded pedestrians. Besides the part-based model, Zhou *et al.* (Zhou and Yuan 2017) present to jointly learn part detectors to exploit part correlations as well as reduce the computational cost. Wang *et al.* (Wang et al. 2018) propose a novel bounding box regression loss to detect pedestrians in the crowd scenes. Zhang *et al.* (Zhang, Yang, and Schiele 2018) propose to utilize channel-wise attention in convnets allowing the network to learn more representative features for different occlusion patterns in one coherent model. Zhang *et al.* (Zhang et al. 2018) design an aggregation loss to enforce proposals to be close and locate compactly to the corresponding objects. Zhou *et al.* (Zhou and Yuan 2018) design a method to detect full body and visible part estimation simultaneously to further estimate occlusion. Although numerous pedestrian detection methods are presented, how to effectively detect each individual pedestrian in crowded scenarios is still one of the most critical issues for pedestrian detectors.

Behind those different methods, there are several datasets (Dalal and Triggs 2005; Ess, Leibe, and Gool 2007; Gerónimo et al. 2007; Overett et al. 2008; Silberstein et al. 2014; Wojek, Walk, and Schiele 2009; Wu and Nevatia 2007) that provide strong support for pedestrian detection in the last decade. The Tsinghua-Daimler Cyclist (TDC) (Li et al. 2016) dataset focuses on cyclists recorded from a vehicle-mounted stereo vision camera, containing a large number of cyclists varying widely in appearance, pose, scale, occlusion and viewpoint. The KITTI (Geiger, Lenz, and Urtasun 2012) dataset focuses on autonomous driving and is collected via a standard station wagon with two high-resolution color and grayscale video cameras, around the mid-size city of Karlsruhe, in rural areas and on highways, up to 15 cars and 30 pedestrians are visible per image. The EuroCity Persons dataset (Braun et al. 2018) provides a large number of highly diverse, accurate and detailed annotations of pedestrians, cyclists and other riders in 31 cities of 12 different European countries.

PedHunter

Figure 1 shows the architecture of PedHunter, which can adopt any existing networks as the backbone. In this work, we use ResNet-50 (He et al. 2016) with 5-level feature pyramid structure as a demonstration. On the basis of FPN (Lin et al. 2017) baseline, we propose three zero-cost components to enhance its occlusion handling ability, *i.e.*, the mask-guided module, the strict classification criterion, and the occlusion-simulated data augmentation. Each component is described below.

Mask-guided Module

It is usually difficult for CNN-based methods to detect occluded pedestrians in crowded scenes, because pedestrians often occlude each other and their features become intertwined after several convolution layers. Thus, more discriminative features of occluded pedestrians need to be learned to hunt these targets in crowded scenes. Compared with other body parts, head is more visible, the occlusion pattern statistics in CityPersons also confirm this. Since head is more visible, people often resort to head to find corresponding pedestrian in crowd scenes. Inspired by this observation, the mask-guided module is designed to utilize the head location as an extra supervision, which assists the network to learn more discriminative features for occluded pedestrians. As shown in Figure 1, this newly added module is in parallel with the existing class&bbox module to predict head masks. In particular, it utilizes four stacked 3×3 conv layers, one deconv layer, one 1×1 conv layer and a per-pixel sigmoid function to predict a binary head mask for each pedestrian proposal. We adopt the box-wise annotation as the segmentation supervision of the proposed module for simplicity. The average binary cross-entropy loss is applied to train this module.

Through the supervision of head segmentation information, the learned features of occluded pedestrians tend to be more discriminative. The feature map visualization of baseline and the proposed PedHunter in Figure 2 also illustrates the enhancement brought by the mask-guided module. Compared to the baseline model, the learned features of PedHunter show stronger response to pedestrians and the distinction of adjacent pedestrians is more obvious. Besides, the proposed module only works in the training phase, so that the detector can keep the same computational cost as the original network during inference.

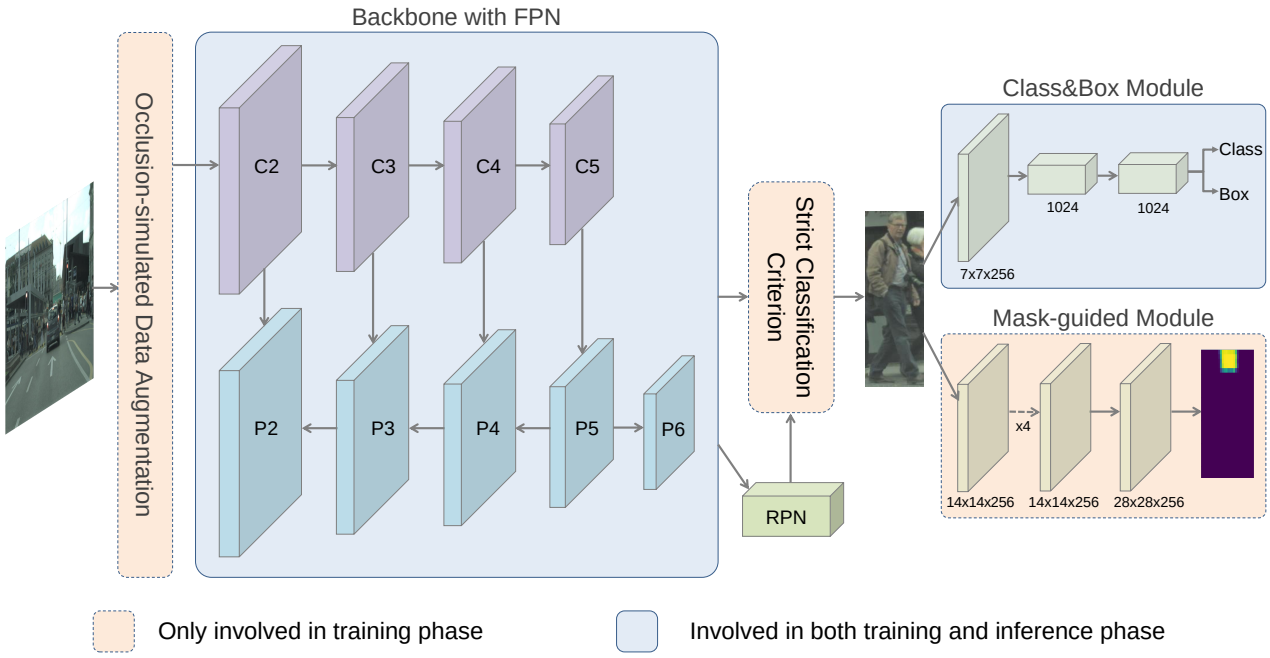


Figure 1: Structure of PedHunter. It contains Backbone with FPN, RPN, Class&Box and Mask-guided Module. The proposed components are shown within orange dotted rectangle. Through Occlusion-simulated Data Augmentation, input images with random occlusion are sent into backbone and then RPN generates candidate bounding boxes. After that, we use strict classification criterion to provide higher-quality positives to subsequent modules. The Class&Box Module performs the second-stage classification and regression, and the Mask-guided Module predicts the associating head mask.

The mask-guided module can be considered as another form of attention mechanism to assist the feature learning. Compared with the external attention guidance method (marked as GA) in (Zhang, Yang, and Schiele 2018), our mask-guided module has the following two advantages: 1) Our method does not use extra dataset, while GA uses the MPII Pose dataset to pre-train part detector. 2) During inference, GA also need the external part detector to generate the attention vector, leading to considerable additional computing overhead. However, our method avoids any additional overhead. Besides, we would like to emphasize that the mask-guided module uses head to **assist the model in learning better features** during training. During inference, even a pedestrian does not have a visible head, the learned better features for this pedestrian can also help the R-CNN branch for better detection performance.

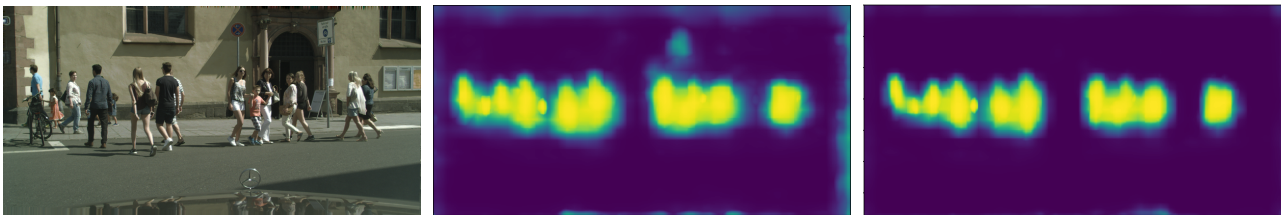
Strict Classification Criterion

Another key problem in occluded pedestrian detection is the false positive between overlapping pedestrians shown in Figure 3(a). These false positives are hard to be suppressed by Non-Maximum Suppression (NMS) because of its sensitivity to the IoU threshold. The main cause for these false positives is the poor quality of positive samples during training R-CNN. As shown in Figure 3(b), RPN often generates some proposals between occluded pedestrians and most of these proposals will be matched as positive samples in R-CNN. During inference, these proposals will be predicted with a high score and thus damage the performance.

To address this issue, we introduce a new strict classification criterion containing the following two steps:

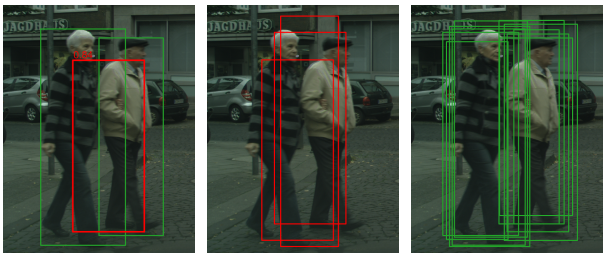
- Increasing IoU threshold for positive samples. To improve the quality of positive samples for R-CNN, we increase the IoU threshold of matching positive samples from 0.5 to 0.7 in the second stage. Only the proposals with higher overlaps with ground truths are treated as positive samples, while those proposals lying in the union of overlapping pedestrians with lower IoU are treated as ignored or negative samples.
- Jittering ground truths. The first step increasing IoU threshold will greatly reduce the number of positive samples, causing a serious class imbalance problem. To solve this issue, we introduce a ground truth jittering operation, which randomly jitters each ground truth 10 times with a small amplitude $[\delta_{x_1},\delta_{y_1},\delta_{x_2},\delta_{y_2}]$, where $\delta_{x_1},\delta_{x_2} \sim Uniform(-0.2w,\ 0.2w)$ and $\delta_{y_1},\delta_{y_2} \sim Uniform(-0.2h,\ 0.2h)$, w and h are width and height of ground truths. Then these jittered boxes are added into proposals to train R-CNN.

As shown in Figure 3(c), the proposed criterion is able to remove those poor-quality positive samples shown in Figure 3(b) and significantly improves the quality of the positive samples. During the inference phase, those poor-quality proposals are more inclined to get lower confidence scores and have a minor impact on the final performance.



(a) Input image (b) Baseline (c) PedHunter

Figure 2: Proposal score feature map visualization. Compared to the baseline, PedHunter shows stronger response to pedestrians, *i.e.*, the color of pedestrian region in (c) is brighter than (b). And the pedestrian boundary on feature map of PedHunter is more clear than baseline, indicating that adjacent pedestrians are more distinguishable in our method.



(a) (b) (c)

Figure 3: (a) A false positive with high score between overlapping pedestrians that is not suppressed by true positives. (b) Several poor-quality positive samples in the R-CNN stage. (c) Improved positive samples under the introduced strict classification criterion.

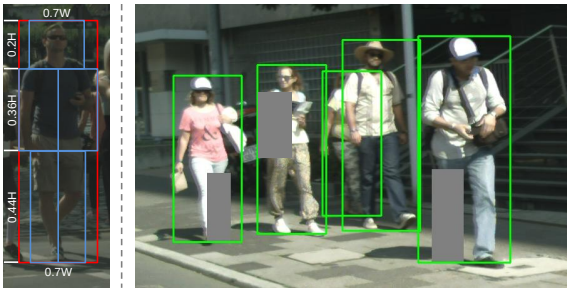


Figure 4: We divide each ground truth into 5 parts (*i.e.*, head, left upper body, right upper body, left leg and right leg). During training, we randomly select one part except the head to add occlusion.

Occlusion-simulated Data Augmentation

Another reason to the relatively poor performance of occluded pedestrian detection is that there is only a small percentage of occlusion cases during training. Taking the CityPersons dataset as an example, more than 50% of the instances only possess less than 20% occlusion ratio. There is fewer occlusion cases in the Caltech-USA dataset.

To diversify the occlusion cases in the training phase, we propose an occlusion-simulated data augmentation to construct a more occlusion-robust model. As shown in Figure 4, we first divide each ground truth into five parts (*i.e.*, head, left upper body, right upper body, left leg, and right leg) with the empirical ratio in (Felzenszwalb et al. 2010), then randomly select one part except the head to occlude with the mean-value of ImageNet (Russakovsky et al. 2015). This data augmentation is used for each ground truth with a probability of 0.5 to ensure randomness.

With the proposed data augmentation, the quantity and pattern of occlusion cases can be significantly enriched during training. Since 50% pedestrians with $1/5$ body are blanked (*i.e.*, only 10% pedestrian regions are blanked and 90% pedestrian regions are natural), thus the detector will not pay excessive attention on added occlusion regions. Besides, similar to that randomly adding some noises on the image can improve the network robustness, this augmentation randomly adds some occlusion parts to improve the occlusion robustness. The improvements brought by this augmentation (stated in Section Model Analyse) can prove the above statement. Additionally, the method also does not bring extra computational cost during the inference phase.

Training and Inference

Anchor Design. At each location of the detection layer, we only associate one specific scale of anchors (*i.e.*, $8S$, where S represents the downsampling factor of the detection layer). They cover the scale range $32 - 512$ pixels across different levels with respect to the network's input image. We use different anchor aspect ratios for different datasets, which is described in the next section.

Sample Matching. Samples are assigned to ground-truth pedestrian boxes using an IoU threshold of θ_p, and to background if their IoU is in $[0, \theta_n)$. If an anchor is unassigned, which may happen with overlap in $[\theta_n, \theta_p)$, it is ignored during the training phase. Based on the proposed strict classification criterion, we set $\theta_n = 0.3$ and $\theta_p = 0.7$ for the RPN stage same as original, and $\theta_n = 0.5$ and $\theta_p = 0.7$ for the R-CNN stage.

Loss Function. The whole network is optimized by $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{box}} + \lambda_2 \mathcal{L}_{\text{mask}}$, where the classification loss $\mathcal{L}_{\text{cls}}$ and the regression loss $\mathcal{L}_{\text{box}}$ are identical as those defined in (Girshick 2015). The mask loss $\mathcal{L}_{\text{mask}}$ is the average binary cross-entropy loss. The loss weight coefficients λ_1 and λ_2 are used to balance different loss terms and we empirically set them as 1 in all experiments.

Optimization. The backbone network is initialized by the ImageNet (Russakovsky et al. 2015) pretrained ResNet-50

Figure 5: (a) An illustrative example of three kinds of annotations: head bounding-box, visible bounding-box and full bounding-box. (b)-(e) Four common scenes in SUR-PED including railway station, intersection, walking street and subway exit.

model. The parameters of newly added layers in the RPN are initialized by the normal distribution method, and the parameters in the class&box module and mask-guided module are initialized by the MSRA normal distribution method. We fine-tune the model using SGD with 0.9 momentum, 0.0001 weight decay. The proposed PedHunter is trained on 16 GTX 1080Ti GPUs with a mini-batch 2 per GPU for CrowdHuman, Caltech-USA and SUR-PED, and the mini-batch size for Citypersons is 1 per GPU. Multi-scale training and testing are not applied to ensure fair comparisons with previous methods. The specific settings of training process for different datasets are described in next section.

Evaluation Metric. The log-average miss rate over 9 points ranging from 10^{-2} to 10^{0} FPPI (*i.e.*, MR^{-2}) is used to evaluate the performance of the detectors (lower score indicates better performance). We report the detection performance for instances in full-body categories.

Inference. During inference, the mask-guided module is not used so that our PedHunter keeps the same computational cost as the baseline. We apply the Non-Maximum Suppression (NMS) with threshold 0.5 to generate the top 100 high confident detections per image as final results.

Datasets

In this section, we first introduce the proposed SUR-PED dataset, then present existing datasets including Caltech-USA, CityPersons and CrowdHuman datasets.

SUR-PED Dataset

The proposed SUR-PED dataset is a benchmark dataset to better evaluate occluded pedestrian detectors in surveillance scenes. It contains $6,000$, $1,000$ and $3,000$ images for training, validation and testing subsets, respectively. The maximum resolution is 1920×1080. Images are crawled using some keywords of surveillance scenes from multiple image search engines including Google, Bing and Baidu. There are totally $162k$ human instances and an average of 16.2 pedestrians per image with various kinds of occlusions. Each human instance is annotated manually with a head bounding-box, a human visible-region bounding-box and a human full-body bounding-box as in CrowdHuman. Different from the bounding boxes with fixed aspect ratio in CityPersons and Caltech, our annotation protocol is more flexible in real world scenarios where have various human poses. Figure 5 illustrates some examples from SUR-PED dataset, with the manner of annotations and the diversity of occlusion. In this paper, we report the detection performance for instances in pedestrian full-body categories. This dataset and a comprehensive evaluation tool will be released further. During training and inference on the proposed dataset, the input images are resized to have a short side of 800 pixels as well as the long edges should be no more than 1333 pixels. Due to the various aspect ratios of ground truths, the anchor aspect ratio setting adopts 0.5, 1 and 2. For the first 13 training epochs, the learning rate is set to 0.04, and we decrease it by a factor of 10 and 100 for another 4 and 3 epochs, respectively.

Caltech-USA Dataset

The Caltech-USA dataset is one of the most popular and challenging datasets for pedestrian detection, which comes from approximately 10 hours 30Hz VGA video recorded by a car traversing the streets in the greater Los Angeles metropolitan area. The training and testing sets contain $42,782$ and $4,024$ frames, respectively. The commonly used $10\times$ training annotations (Zhang et al. 2016b) of Caltech-USA are refined automatically with only $16,376$ poor-quality instances in the training set. In our previous work, we re-annotate the dataset manually following the labeling rule and method in CrowdHuman as the extended version.

Table 1: Ablation experiments using MR^{-2} (lower score indicates better performance). For CityPersons and Caltech datasets, we only report results on the Reasonable set.

Method	Strict Criterion	Mask-guided	Occlusion -simulated	Backbone	CityPersons	Caltech	CrowdHuman	SUR-PED
Baseline				ResNet-50	11.67	3.26	46.8	57.9
PedHunter	√			ResNet-50	10.59	2.91	45.2	56.7
	√	√		ResNet-50	9.62	2.64	43.6	55.7
	√	√	√	ResNet-50	8.32	2.31	39.5	53.6

We train the proposed method using $2\times$ scale of the image size with 2.44 as the anchor aspect ratio. The initial learning rate is 0.04 for the first 4 epochs, and is reduced by 10 and 100 times for another 2 and 1 epochs.

CityPersons Dataset

The CityPersons dataset is recorded across 18 different cities in Germany with 3 different seasons and various weather conditions. The dataset includes $5,000$ images ($2,975$ for training, 500 for validation, and $1,525$ for testing) with $\sim 35,000$ manually annotated persons plus $\sim 13,000$ ignored annotations. In our previous work, we additionally label the corresponding head bounding box for each annotated pedestrian instance as the extended version. The proposed PedHunter detector is trained on the training set and evaluated on the validation set. We enlarge input images by 1.3 times and only use 2.44 anchor aspect ratio for training. The initial learning rate is set to 0.02 for the first 26 epochs, and is decreased to 0.002 and 0.0002 for another 9 and 5 epochs.

CrowdHuman Dataset

The CrowdHuman dataset is a benchmark dataset to evaluate pedestrian detectors in crowd scenarios. It is divided into training ($15,000$ images), validation ($4,370$ images) and testing ($5,000$ images) subsets. In particular, there are totally $470k$ human instances from the training and validation subsets, and 22.6 persons per image with various kinds of occlusions. Each human instance is annotated with a head bounding-box, a human visible-region bounding-box and a human full-body bounding-box. Since the online evaluation server for the testing subset is not available until now, we train the proposed models on the CrowdHuman training subset and evaluate on the validation subset. During training, the input images are resized so that their short edges are at 800 pixels while the long edges should be no more than 1333 pixels at the same time. The anchor aspect ratios are set to 0.5, 1 and 2 for CrowdHuman dataset. We train PedHunter with the initial learning rate 0.04 for the first 16 epochs, and decay it by 10 and 100 times for another 6 and 3 epochs.

Experiments

In this section, we first analyze the effectiveness of the proposed method, then evaluate the final model on common pedestrian detection benchmark datasets.

Model Analyse

To sufficiently verify the effectiveness of the proposed components, we construct ablation study on all four datasets including CrowdHuman, CityPersons, Caltech and SUR-PED. We first construct a baseline detector based on FPN (Lin et al. 2017) with ResNet-50 (He et al. 2016). The performance of the baseline model is shown in Table 1. For the CityPersons dataset, it obtains 11.67% MR^{-2} on the Reasonable set, outperforming the adapted FasterRCNN baseline in CityPersons by 1.13%. For the Caltech dataset, it achieves 3.26% MR^{-2} on the Reasonable set, which already surpasses all the state-of-the-art methods. For the CrowdHuman dataset, our implemented FPN baseline gets 46.8% MR^{-2} on the validation set, which is 3.62% better than the reported baseline in (Shao et al. 2018). Thus, our baseline models are strong enough to verify the effectiveness of the proposed components. From the baseline model, we gradually apply the proposed strict classification criterion, mask-guided module and occlusion-simulated data augmentation method to verify their effectiveness. The parameter setting of all detectors in both training and testing is consistent for the fair comparison.

Strict Classification Criterion We first apply strict classification criterion onto baseline detectors to demonstrate its effectiveness. Comparing the detection results in Table 1, we find that using the newly proposed strict classification criterion is able to reduce the MR^{-2} by 1.08% from 11.67% to 10.59% on the CityPersons dataset, by 0.35% from 3.26% to 2.91% on the Caltech dataset, by 1.6% from 46.8% to 45.2% on the CrowdHuman dataset, and by 1.2% from 57.9% to 56.7% on the proposed dataset. We also provide visualization comparison of positive samples between whether applying the strict classification criterion or not in Figure 3. It is obvious that the strict classification criterion effectively improves the quality of the positive samples in the R-CNN stage. Both performance improvements on four datasets and the visualization comparison demonstrate that strict classification criterion is effective for detecting the pedestrians in crowded scenes.

Mask-guided Module To validate the effectiveness of the mask-guided module, we add it after applying strict classification criterion. The ablation results are shown in Table 1. Adding mask-guided module decreases the MR^{-2} from 10.59% to 9.62% with 0.97% improvement for the CityPersons dataset, from 2.91% to 2.64% with 0.27% improvement for the Caltech dataset, from 45.2% to 43.6% with

Table 2: MR^{-2} performance on heavy occlusion subset.

Strict Criterion	Mask-guided	Occlusion -simulated	CityPersons	Caltech-USA
			49.74	53.26
√			47.44	51.10
√	√		46.33	48.64
√	√	√	**43.53**	**45.31**
RepLoss			55.30	63.36
OR-CNN			51.30	69.57

1.6% improvement for the CrowdHuman dataset, and from 56.7% to 55.7% with 1.0% improvement for the proposed dataset. These improvements fully demonstrate the effectiveness of the proposed mask-guided module in occluded pedestrian detection. Meanwhile, we also present visualization comparison of feature maps between whether applying mask-guided module or not in Figure 2. It is relatively obvious that mask-guided module contributes to more aggregate and strong response to pedestrian instances.

Occlusion-simulated Data Augmentation We propose the occlusion-simulated data augmentation to enrich the pattern and quantity of occlusion cases during training for stronger occlusion robustness. As shown in Table 1, when we add this proposed augmentation strategy after the previous two improvements, the MR^{-2} is further reduced by 1.3% (8.32% *vs.* 9.62%) on the CityPersons dataset, by 0.33% (2.31% *vs.* 2.64%) on the Caltech dataset, by 4.1% (39.5%*vs.* 43.6%) on the CrowdHuman dataset, and by 2.1% (53.6% *vs.* 55.7%) on the proposed SUR-PED dataset. These comprehensive improvements indicate the effectiveness of the presented occlusion-simulated data augmentation for occluded pedestrian detection in a crowd.

Occlusion Subset Performance We also report the results on the heavy occlusion subset of CityPersons and Caltech-USA datasets in Table 2. After gradually applying the proposed three contributions, the MR^{-2} performances are significantly improved by 6.21% and 8.05%, respectively. Besides, our method shows a large margin over state-of-the-art methods, *i.e.*, RepLoss and OR-CNN. These results on heavy occlusion subset demonstrate the effectiveness of the proposed PedHunter method.

Benchmark Evaluation

CityPersons. We compare PedHunter with TLL (MRF) (Song et al. 2018), Adapted FasterRCNN (Zhang, Benenson, and Schiele 2017), ALFNet (Liu et al. 2018), Repulsion Loss (Wang et al. 2018), PODE+RPN (Zhou and Yuan 2018) and OR-CNN (Zhang et al. 2018) on the CityPersons validation subset in Table 3. Similar with previous works, we evaluate the final model on the Reasonable subset of the CityPersons dataset. The proposed PedHunter method surpasses all published methods and reduces the MR^{-2} score of state-of-the-art results from 11.0% to 8.32% with 2.68% improvement compared with the second best method (Zhang et al. 2018), demonstrating the superiority of the proposed method in pedestrian detection.

Table 3: MR^{-2} performance on the CityPersons validation set. Scale indicates the enlarge number of original images.

Method	Backbone	Scale	*Reasonable*
TLL (MRF)	ResNet-50	-	14.40
Adapted FasterRCNN	VGG-16	×1.3	12.97
ALFNet	VGG-16	×1	12.00
Repulsion Loss	ResNet-50	×1.3	11.60
PODE+RPN	VGG-16	-	11.24
OR-CNN	VGG-16	×1.3	11.00
PedHunter	ResNet-50	×**1.3**	**8.32**

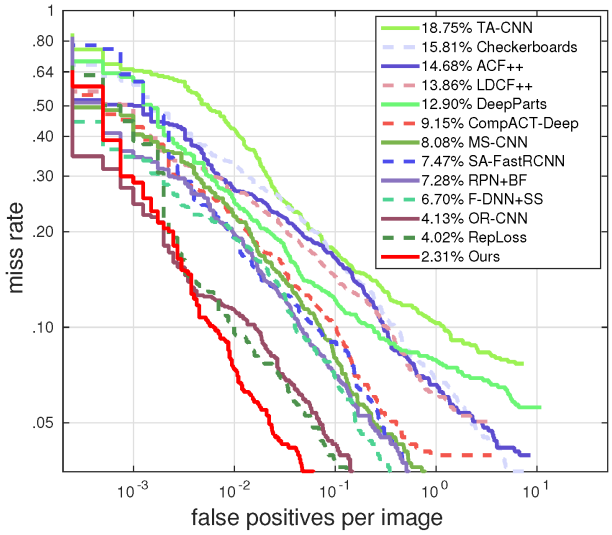


Figure 6: MR^{-2} scores of different state-of-the-art methods on the Caltech-USA dataset.

Caltech-USA. Figure 6 shows the comparison of the PedHunter method with other state-of-the-art methods on the Caltech-USA testing set. All the reported results are evaluated on the widely-used Reasonable subset, which only contains the pedestrians with at least 50 pixels tall and occlusion ratio less than 35%. The proposed method outperforms all other state-of-the-art methods by producing 2.31% MR^{-2}.

Conclusion

This paper presents an efficient method to improve the occluded pedestrian detection accuracy in crowded scenes. Specifically, a mask-guided module is designed to enhance the representation and discrimination of features. Meanwhile, a strict classification criterion is introduced to eliminate common false positives in crowded scenes. Moreover, an occlusion-simulated data augmentation method is proposed to improve the robustness of network against occlusion. Besides, we collect a new occluded pedestrian detection benchmark dataset in surveillance scenes. Consequently, we achieve state-of-the-art performances on common pedestrian detection datasets. The proposed dataset, source codes and trained models will be public to facilitate further studies of occluded pedestrian detection.

References

Angelova, A.; Krizhevsky, A.; Vanhoucke, V.; Ogale, A. S.; and Ferguson, D. 2015. Real-time pedestrian detection with deep network cascades. In *BMVC*.

Braun, M.; Krebs, S.; Flohr, F.; and Gavrila, D. M. 2018. The eurocity persons dataset: A novel benchmark for object detection. *CoRR*.

Brazil, G.; Yin, X.; and Liu, X. 2017. Illuminating pedestrians via simultaneous detection and segmentation. In *ICCV*.

Cai, Z.; Saberian, M. J.; and Vasconcelos, N. 2015. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*.

Costea, A. D., and Nedevschi, S. 2016. Semantic channels for fast pedestrian detection. In *CVPR*.

Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2009. Pedestrian detection: A benchmark. In *CVPR*.

Du, X.; El-Khamy, M.; Lee, J.; and Davis, L. S. 2017. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*.

Ess, A.; Leibe, B.; and Gool, L. J. V. 2007. Depth and appearance for mobile scene analysis. In *ICCV*.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D. A.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *TPAMI*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*.

Gerónimo, D.; Sappa, A.; López, A.; and Ponsa, D. 2007. Adaptive image sampling and windows classification for on-board pedestrian detection. In *ICVS*.

Girshick, R. B. 2015. Fast R-CNN. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hosang, J. H.; Omran, M.; Benenson, R.; and Schiele, B. 2015. Taking a deeper look at pedestrians. In *CVPR*.

Li, X.; Flohr, F.; Yang, Y.; Xiong, H.; Braun, M.; Pan, S.; Li, K.; and Gavrila, D. M. 2016. A new benchmark for vision-based cyclist detection. In *IVS*.

Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; and Yan, S. 2018. Scale-aware fast R-CNN for pedestrian detection. *TMM*.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2017. Feature pyramid networks for object detection. In *CVPR*.

Lin, C.; Lu, J.; Wang, G.; and Zhou, J. 2018. Graininess-aware deep feature learning for pedestrian detection. In *ECCV*.

Liu, W.; Liao, S.; Hu, W.; Liang, X.; and Chen, X. 2018. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *ECCV*.

Mao, J.; Xiao, T.; Jiang, Y.; and Cao, Z. 2017. What can help pedestrian detection? In *CVPR*.

Noh, J.; Lee, S.; Kim, B.; and Kim, G. 2018. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *CVPR*.

Overett, G.; Petersson, L.; Brewer, N.; Andersson, L.; and Pettersson, N. 2008. A new pedestrian dataset for supervised learning. In *IVS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *IJCV*.

Sermanet, P.; Kavukcuoglu, K.; Chintala, S.; and LeCun, Y. 2013. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. Crowdhuman: A benchmark for detecting human in a crowd. *CoRR*.

Silberstein, S.; Levi, D.; Kogan, V.; and Gazit, R. 2014. Vision-based pedestrian detection for rear-view cameras. In *IVS*.

Song, T.; Sun, L.; Xie, D.; Sun, H.; and Pu, S. 2018. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *ECCV*.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015a. Deep learning strong parts for pedestrian detection. In *ICCV*.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015b. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*.

Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; and Shen, C. 2018. Repulsion loss: Detecting pedestrians in a crowd. In *CVPR*.

Wojek, C.; Walk, S.; and Schiele, B. 2009. Multi-cue onboard pedestrian detection. In *CVPR Workshop*.

Wu, B., and Nevatia, R. 2007. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *ICCV*.

Yang, B.; Yan, J.; Lei, Z.; and Li, S. Z. 2015. Convolutional channel features. In *ICCV*.

Zhang, S.; Benenson, R.; and Schiele, B. 2015. Filtered channel features for pedestrian detection. In *CVPR*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

Zhang, L.; Lin, L.; Liang, X.; and He, K. 2016a. Is faster R-CNN doing well for pedestrian detection? In *ECCV*.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J. H.; and Schiele, B. 2016b. How far are we from solving pedestrian detection? In *CVPR*.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *ECCV*.

Zhang, S.; Yang, J.; and Schiele, B. 2018. Occluded pedestrian detection through guided attention in CNNs. In *CVPR*.

Zhou, C., and Yuan, J. 2017. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*.

Zhou, C., and Yuan, J. 2018. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*.