EasyChair Preprint

№ 3783

Semantic Fusion-based Pedestrian Detection for Supporting Autonomous Vehicles

Mingzhi Sha and Azzedine Boukerche

July 7, 2020

Semantic Fusion-based Pedestrian Detection for Supporting Autonomous Vehicles

Mingzhi Sha, Azzedine Boukerche
PARADISE Research Laboratory, EECS, PARADISE Research Laboratory, EECS, University of Ottawa, Canada
Emails: msha096@uottawa.ca, boukerch@eecs.uottawa.ca

***Abstract*—To increase traffic safety and transportation efficiency, adopting intelligent transportation systems (ITS) has become a trend. As an important component of ITS, one essential task of autonomous vehicles is to detect pedestrians accurately, which is of great significance for improving traffic safety and building a smart city. In this paper, we propose an anchor-free pedestrian detection model named Bi-Center Network (BCNet) by fusing the full body center and visible part center for each pedestrian. Experimental results show that the performance of pedestrian detection can be improved with a strengthened heatmap, which combines the full body with the visible part semantic. We compare our BCNet with state-of-the-art models on the CityPersons dataset and the ETH dataset, which shows that our approach is effective. Compared to the backbone model, our BCNet improves the detection accuracy by 1.2% on the Reasonable setup and Partial Setup of the CityPersons dataset.**

***Index Terms*—Intelligent transportation system, autonomous vehicle, pedestrian detection, convolutional neural network.**

I. Introduction

The transportation system is one essential foundation of the modern city and becomes much more complicated with the progress of society. To reduce the transportation burden, arrange transportation reasonably, and improve safety and quality of human life, the concept of intelligent transportation systems (ITS) was proposed. ITS is an advanced system that enables all the transport participants to collaborate, exchange information, and understand surroundings, thereby efficiently improving transportation safety and playing an important role in building smart cities. Pedestrian detection is one of the fundamental tasks for supporting autonomous vehicles and other essential components of ITS, and the detection information can be shared among autonomous vehicles by taking advantage of the vehicular networks. In recent years, the widespread application of pedestrian detection has enabled more researchers to contribute to this area [1]–[4].

As mentioned in [3] and [4], the occlusion is very common in real scenarios, and accordingly, many researchers have worked on occlusion handling. However, we have some reasons to point out that heavy occlusion handling (i.e., the pedestrian is occluded more than 35% of the full body, as defined in [3]) in pedestrian detection is not the most urgent task for supporting autonomous vehicles.

This work is partially supported by NSERC-SPG, NSERC-DISCOVERY, Canada Research Chairs Program, NSERC-CREATE TRANSIT Funds.

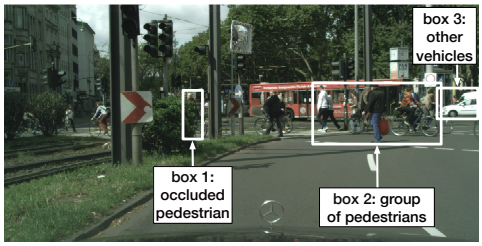


Fig. 1. The pedestrian behind the bush (in box 1) is occluded in the perspective of the current dash camera, but is not occluded in the perspective of other vehicles (in box 3) on another roadside. The group of pedestrians are drawn in box 2.

Firstly, there are many studies on occlusion handling, but the detection accuracy of heavily occluded pedestrians is much lower than other types. Secondly, as mentioned in [5], around 48.8% of pedestrians are occluded by other pedestrians in the CityPersons dataset [4], but autonomous vehicles are not required to predict how many pedestrians are in the crowd. No matter how many pedestrians are in box 2 of Fig. 1, the autonomous vehicle should avoid hurting them. Lastly, autonomous vehicles can exchange information with the help of ITS. There are many studies on coping with the content delivery issue in vehicular networks, e.g., [6]–[8], which enable the images of the same pedestrian captured by different vehicles to be shared with other vehicles. In shared images from different roadsides and observation points, there is likely to be at least one image for each pedestrian that is slightly occluded or even unoccluded. In Fig. 1, it is very dangerous that the autonomous vehicle may hurt the occluded pedestrian behind the bush. However, with the help of information exchange, the white vehicle will send the detection result of this pedestrian to all other vehicles in this area so that every autonomous vehicle will notice this pedestrian even if it cannot observe that pedestrian from its perspective.

In this paper, we focus on improving the performance of pedestrian detection under reasonable occlusion. The proposed Bi-Center Network (BCNet) is an anchor-free network, and our work has the following contributions:

1) We utilize the visible part semantic of each pedestrian and fuse with full body semantic to obtain the enriched feature of each pedestrian.
2) We do an ablation study to find out how to balance

the hyper-parameters of the full body semantic and the visible part semantic, and we visualize the enhanced response of the fused center keypoint on the heatmap.

In the rest of this paper, we will first introduce machine learning methods development and classify the existing neural network-based detectors from two aspects in Section II. In Section III, we will illustrate the details and innovations of our model. Following in Section IV, we conduct experiments on the CityPersons dataset [4] and the ETH dataset [9], implementation details and evaluation results will be given. Lastly, the conclusion of our experiments and future work plans will be shown in Section V.

II. Related Work

Back to more than a decade ago, object detection relied mainly on traditional machine learning methods. One of the essential differences between traditional machine learning methods and deep learning methods is that the former use predefined feature descriptors (e.g., SIFT [10], HOG [11], and Haar [12]) to extract features, while the latter learn and extract features from the datasets by using neural networks in the training phase. With the development of computing hardware, the computing ability is largely increased. In 2012, Krizhevsky et al. proposed AlexNet [13], and since then, the convolutional neural network is getting attention again and blossoming. In this section, we will classify the existing deep learning-based detectors from two aspects, and analyze the pros and cons for each type of detectors.

A. One-stage detectors vs. two-stage detectors

Detectors can be roughly divided into two branches: two-stage detectors (TSDs) and one-stage detectors (OSDs). The most representative TSDs are R-CNN [14], Fast R-CNN [15], and Faster R-CNN [16] family. YOLO [17] and SSD [18] are two classic OSDs. The main difference is that TSDs have a proposal generation phase before generating the final bounding boxes (bbox), while OSDs do not. Correspondingly, TSDs are usually slower than OSDs but with higher precision. One reason behind OSDs' unsatisfactory precision is that OSD has to classify approximately 100k candidate bboxes, while TSD only needs to process around 2000 candidate bboxes with the first stage's help. Consequently, most of the candidates that OSD needs to process are hard negative samples, which can overwhelm the entire training phase. In essence, this is the imbalance between foreground and background classes. To cope with class imbalance, Lin et al. designed the Focal Loss [19] to re-weight positive samples and negative samples in the training phase. Accordingly, the detection accuracy of OSDs has been dramatically enhanced.

B. Anchor-based detectors vs. anchor-free detectors

Anchor boxes are candidates for the region proposals generated by TSDs in the first stage, and candidates for the final bboxes of OSDs. Since Faster R-CNN [16] came out, more pedestrian detectors have tended to use the anchor-based framework with predefined anchor boxes, such as RepLoss [5], OR-CNN [20], and Bi-box [21]. RepLoss [5] enforces the predicted bbox far away from other ground truth pedestrians and their designated proposals. However, one drawback of RepLoss is that it does not consider the overlapping pedestrians in crowded scenes, as indicated in [20]. Bi-box [21] is capable of detecting pedestrians and estimating occlusion simultaneously by applying parallel branches in the detection head. Compared to anchor-based detectors, anchor-free detectors are more flexible because they do not adopt the hyper-parameters such as scales and ratios of anchor boxes. To achieve anchor-free detectors, the use of alternative annotations is a primary method. TLL [22] is an anchor-free detector that predicts the somatic topological line of each pedestrian. CornerNet [23] achieves an anchor-free framework by detecting two corner keypoints of each object. Inspired by CornerNet [23], Duan et al. proposed CenterNet [24], which utilizes the center keypoint together with top-left and bottom-right corner keypoints. By predicting the center keypoint and the corresponding scale of each pedestrian, Center and scale prediction-based detector (CSP) [25] greatly improved the detection accuracy on the CityPersons [4] dataset.

Based on the above analyses, our proposed BCNet will take advantage of the OSD because of its fast speed and high accuracy. Meanwhile, we will use the center keypoint to locate the pedestrian in that the center keypoint has the internal feature of the pedestrian.

III. Proposed Approach

One main feature of our BCNet is to predict two heatmaps: one for the full body center keypoints and another for the visible part center keypoints. These two heatmaps are fused to generate the enhanced feature maps. Our model underlines the utility of the visible part semantic.

A. Model overview

The model proposed in this paper takes CSP [25] as the backbone. One drawback of CSP is that it does not take advantage of the visible part feature for each pedestrian. In the proposed network, we introduce a heatmap for the center keypoint of the visible part. The model architecture is shown in Fig. 2. We use ResNet-50 [26] structure in ConvNet to extract different levels of features. The feature map from ConvNet's deeper layers has a lower resolution but a higher semantic level. To take advantage of the high resolution and high semantic feature, we extract multi-scale feature maps from conv2_x layer, conv3_x layer, conv4_x layer, and conv5_x layer in ResNet-50. Before concatenating together, we rescale four feature maps to the same size by using the transposed convolution layers. The generated final feature map is of size H/r$\times$W/r, where H and W are the height and width of the input image, and r = 4 is the downsampling ratio suggested in [27]. In our model, after reducing the feature map channel from 1024 to 256 by a 3$\times$3 Conv layer, our detector has four parallel branches. These four branches are processed by four separate 1$\times$1 Conv layers and the parameters of the four subnets are not shared. The outputs of the full body center keypoint branch

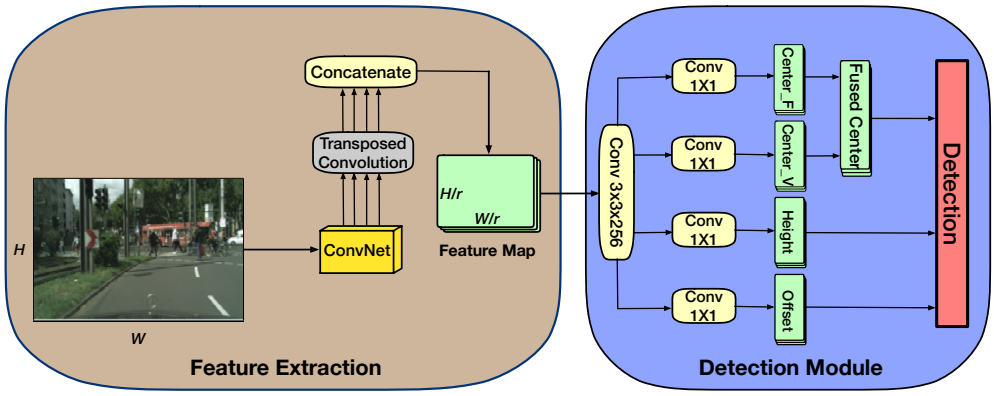


Fig. 2. The architecture of BCNet. We take ResNet-50 [26] as backbone in ConvNet. Center_F denotes the full body center keypoint prediction and Center_V denotes the visible part center keypoint prediction.

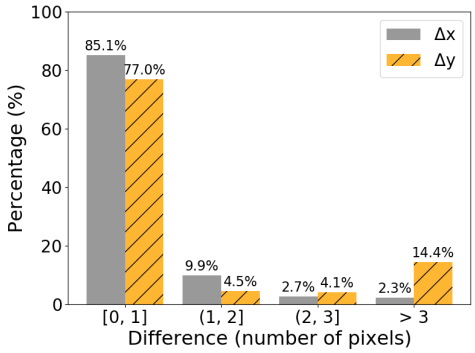


Fig. 3. Distribution of distances between the visible part center keypoint and the full body center keypoint for each pedestrian in the CityPersons training set in the resolution of 256×512. Δx and Δy denote horizontal distance and vertical distance, respectively.

and visible part center keypoint branch are predicted heatmaps, on which we can locate pedestrians through the high response points. To generate the detection result, we first fuse center keypoints on these two predicted heatmaps and then use the score threshold to filter out the high response points on the fused heatmap. We fine-tune the location by the predicted offset. With the predicted height, we can multiply the height by the aspect ratio to get the pedestrian width. In our work, we use a fixed aspect ratio of 0.41 as defined in [1].

B. Pedestrian detection with semantic fusion

In Fig. 3, we analyze the distribution of distances between the visible part center keypoint and the full body center keypoint for each annotated pedestrian in the CityPersons [4] training set. 19654 annotated objects are labeled as pedestrians. In the resolution of the final feature maps that have been downsampled with factor r = 4, around 85% and 77% of the pedestrians have a distance of no larger than one pixel between their full body centers and visible part centers horizontally and vertically, respectively. For the pedestrians under reasonable occlusion (visibility ratio ≥ 0.65, as defined in [4]), the visible part center keypoint and the full body center keypoint should be very close and even identical. For the heavily occluded (occlusion ratio ≥ 0.35) pedestrians, two center keypoints are usually far apart spatially. Thereby, we combine the location sensitive confidence scores obtained from two center keypoint heatmaps as in:

$$\alpha Score_F + \beta Score_V = Score, \tag{1}$$

where $Score_F$ and $Score_V$ represent the confidence score of the full body center keypoint and the visible part center keypoint, respectively. α and β are weighting factors $\in$ [0, 1]. The score in Eq. (1) is determined based on location on the heatmap, which means the confidence scores at different locations will not affect each other. More details about choosing hyper-parameters α and β will be given in Section IV-B1. Notably, even if the confidence scores are not combined, the visible part center keypoint branch can naturally benefit the model to converge.

C. Loss function

We formulate two center keypoints prediction branches as classification problems. L_{cls_f} and L_{cls_v} denote the loss of the full body center keypoint branch and the visible part center keypoint branch, respectively. The ground truth heatmaps are generated by the 2D Gaussian function introduced in [25].

In the visible part center keypoint branch, y_{ij} represents the $Score_V$ on the ground truth heatmap, and p_{ij} represents the $\widehat{Score_V}$ on the predicted heatmap. Both y_{ij} and $p_{ij} \in$ [0, 1]. p_{ij} is the predicted probability to indicate if there is a center keypoint of the visible part pedestrian at the location (i, j). Similar to [23] and [25], we modify the Focal Loss function introduced in [19], which is given by Eq. (2). In this equation, N is the number of annotated pedestrians in the image. H_0 and W_0 represent the height and width of the input after downsampling, respectively. γ and δ are focusing hyper-parameters, and in the experiment we set γ = 2, δ = 4 as suggested in [23]. The same classification loss function in Eq. (2) is used to calculate L_{cls_f} in the full body center keypoint prediction branch.

$$L_{cls_v} = -\frac{1}{N}\sum_{i=1}^{H_0}\sum_{j=1}^{W_0}\begin{cases}(1-p_{ij})^{\gamma}\log(p_{ij}) & \text{if } y_{ij}=1\ ,\\ (1-y_{ij})^{\delta}(p_{ij})^{\gamma}log(1-p_{ij}) & \text{otherwise}\ .\end{cases} \tag{2}$$

In the height and offset prediction branches, we formulate them as regression problems. We use the smooth L1 Loss function [15] at the center of full body locations, as in:

$$L_{scale} = \frac{1}{N}\sum_{k=1}^{N} SmoothL1Loss(h_k, \hat{h_k}), \tag{3}$$

$$L_{offset} = \frac{1}{N}\sum_{k=1}^{N} SmoothL1Loss(o_k, \hat{o_k}), \tag{4}$$

where h_k, $\hat{h_k}$, o_k, and $\hat{o_k}$ are the ground truth height, predicted height, ground truth offset, and predicted offset of pedestrian k, respectively.

The final loss function to be optimized in the training phase is added up as in:

$$Loss = \lambda_f L_{cls_f} + \lambda_v L_{cls_v} + \lambda_s L_{scale} + \lambda_o L_{offset}, \tag{5}$$

where λ_f, λ_v, λ_s, and λ_o are weighting factors for the losses in each branch, and we experimentally set them to 0.01, 0.01, 1, and 0.1, respectively.

IV. Experiments

In this section we will demonstrate how the hyper-parameters α and β in Eq. (1) affect the performance, and we conduct experiments on the CityPersons dataset [4] and the ETH dataset [9]. We adopt the unified evaluation metric MR^{-2} (the lower the better) introduced in [1], which is the mean value of nine derived miss rates with the corresponding FPPIs (false positive per image) evenly located in $[10^{-2}, 10^{0}]$ within the log-space. The experimental results shown in Table II, Fig. 4, and Fig. 6 are all evaluated by MR^{-2}.

A. Dataset and experimental setup

The CityPersons [4] dataset provides annotations of the visible part for each pedestrian; it has large resolution images (1024 $\times$2048); it covers multiple seasons and countries. The ETH [9] dataset has a considerable number of pedestrians per image, but the resolution of images (640 $\times$480) is much lower. These properties make them representative, and we choose them to conduct our experiments.

We train our model on the CityPersons [4] training set and test on the validation set, with 2975 and 500 images, respectively. We use the model that trained on the CityPersons dataset [4] to directly test on the ETH [9] dataset without training or fine-tuning. The reason why we test our model on low-resolution images after training on high-resolution images is to avoid overfitting and test generalization of our model.

Our BCNet is trained on single Nvidia GeForce GTX 1080 Ti GPU with the mini-batch size of 2 for the CityPersons dataset [4]. The proposed model is implemented in Python 2.7 and PyTorch 1.2.0.

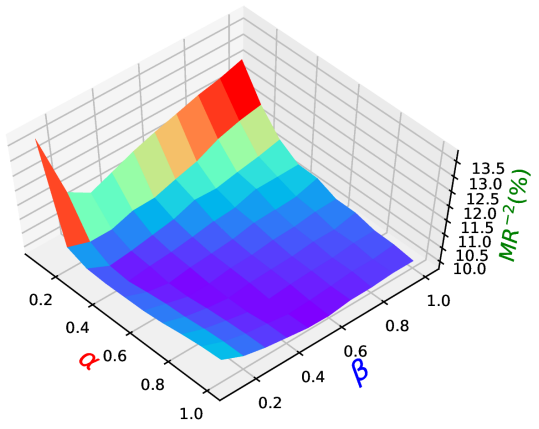


Fig. 4. Experiments of varying α and β on the Reasonable setup of the CityPersons dataset, evaluated by MR^{-2}.

B. Experiments on the CityPersons dataset

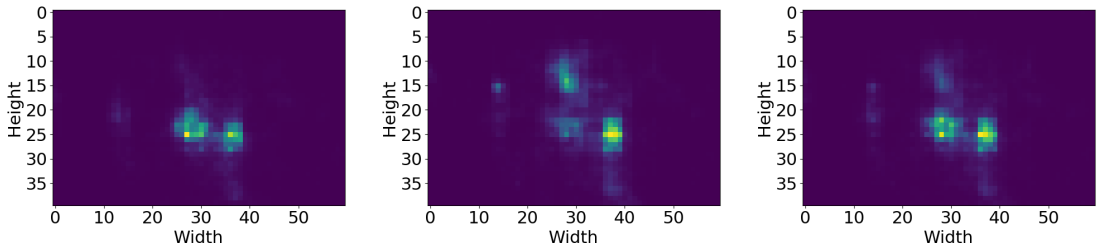
In our experiments, we use a total of seven setups (Reasonable, Bare, Partial, Heavy, Small, Medium, and Large) to evaluate MR^{-2} and the configurations for each setup are shown in Table I. For example, the pedestrian samples on the Reasonable setup are at least 50 pixels in height and visible at least 65% of the full body.

1) Ablation study: In Eq. (1), we introduced two hyper-parameters α and β, which are the weights for the confidence scores of the full body center keypoint and visible part center keypoint, respectively. By changing α and β, the final score will be fused by different ratios of two confidence scores. We do the ablation study on the Reasonable setup of the CityPersons dataset, of which the occlusion ratio is under 0.35. The results of how α and β can affect the performance on the Reasonable setup are shown in Fig. 4, which is similar when applied to other setups in Table I. Obviously, there is a region that MR^{-2} is low and stable, and this region appears when the ratio of α and β is around 2:1, where $\alpha \in [0.4, 1]$ and $\beta \in [0.2, 0.6]$. These (α, β) combinations work well, and the performance is promising and stable. To simplify, we use $\alpha = 1$ and $\beta = 0.5$ to conduct our experiments and evaluations.

The predicted heatmaps are visualized in Fig. 5(a) and Fig. 5(b). Fig. 5(c) is the final enhanced heatmap by fusing Fig. 5(a) and Fig. 5(b) with Eq. (1), where the weight α = 1 and β = 0.5. The bright spot on the heatmap indicates the location where the predicted confidence score is higher. In other words, the bright point is the predicted center of the full body and visible part for each pedestrian. In Fig. 5(c) the fused body center is brighter than the one before integration, which indicates it has a stronger response and higher confidence

TABLE I
EVALUATION SETUPS OF THE CITYPERSONS DATASET [4]

	Reasonable	***Bare***	***Partial***	***Heavy***	***Small***	***Medium***	***Large***
Height (in pixels)	[50, $+\infty$]	[50, $+\infty$]	[50, $+\infty$]	[50, $+\infty$]	[50, 75]	[75, 100]	[100, $+\infty$]
Visibility ratio	[0.65, 1]	[0.9, 1]	[0.65, 0.9]	[0, 0.65]	[0.65, 1]	[0.65, 1]	[0.65, 1]



(a) The heatmap of full body centers. (b) The heatmap of visible part centers. (c) The heatmap after score fusion.

Fig. 5. The visualized heatmaps are cropped into size 40×60 from 256×512 after downsampling with factor r = 4. Width and height are in pixels.

TABLE II
EXPERIMENTAL RESULTS ON THE CITYPERSONS VALIDATION SET. THE RESULTS IN BOLDFACE INDICATE THE BEST PERFORMANCE.

	Reasonable(%)	***Bare(%)***	***Partial(%)***	***Heavy(%)***	***Small(%)***	***Medium(%)***	***Large(%)***
Faster R-CNN with Semantic [4]	14.8	-	-	-	22.6	6.7	8.0
RepLoss [5]	13.2	7.6	16.8	56.9	-	-	-
OR-CNN [20]	12.8	6.7	15.3	55.7	-	-	-
ALFNet [28]	12.0	8.4	11.4	51.9	19.0	5.7	6.6
RFBNet (with adaptive-NMS) [29]	12.7	7.6	11.7	51.9	-	-	-
CSP (with offset) [25]	11.0	7.3	10.4	**49.3**	16.0	3.7	6.5
BCNet (ours)	**9.8**	**5.8**	**9.2**	53.3	**13.0**	**3.3**	**6.1**

score. In the meanwhile, the center of the visible part can be kept, which helps to enrich the semantic information for each pedestrian. We observe that the confidence scores of unoccluded and slightly occluded pedestrians are enhanced; while when the pedestrians are under heavy occlusion, the confidence scores are not dramatically affected since two center keypoints are not close spatially.

2) Evaluation and analysis: During the training phase on the CityPersons dataset, we initial our model weights with backbone ResNet-50 [26] which was pre-trained on ImageNet [30], and we use Adam [31] to minimize the loss in Eq. (5) for a total of 100 epochs. We use a learning rate of 5×10^{-5} for the first 50 epochs and 2×10^{-5} for the last 50 epochs. The locations where the scores above 0.1 on the fused heatmap are kept, and the candidate bbox will be generated based on the retained center keypoint, predicted height, and offset. The candidate bbox will then be applied with non-maximum suppression (NMS) at a threshold of 0.5. We adopted the same standard data augmentation techniques as demonstrated in [25] to increase the diversity of data and help reduce the overfitting for the dataset.

We test our BCNet on the CityPersons [4] validation set with the batch size of 1, and it takes 0.32s to infer one image with the original resolution on a single Nvidia GeForce GTX 1080 Ti GPU. We compare the results with other state-of-the-art models in Table II by using the same comparison method with other works like [20], [5], and [25]. Evaluation results are cited from the published works. The results not provided in the original works are indicated by the symbol '-' in Table II.

We demonstrate the performance of state-of-the-art models in Table II. Compared to the baseline model CSP [25] with about 40M parameters, our BCNet only introduces 257 additional parameters to achieve such a significant improvement; our model achieves 3.0% better MR^{-2} on the Small setup, 1.5% better MR^{-2} on the Bare setup, and 1.2% better MR^{-2} on the Reasonable setup and Partial setup. Obviously, our BCNet achieves promising performance on almost all setups. Notably, our model gained huge success on the Small setup, with the height of each pedestrian varying in $[50, 75]$, which will help to detect pedestrians in farther distances and could help autonomous vehicles have a longer time to react. Although we did not consider heavy occlusion processing, our model still achieves a moderate performance on the Heavy setup and even beats OR-CNN and RepLoss which are designed to handle occlusion. The overall promising performance of our model lies in the usage and combination of semantic information, which will contribute to vehicular networks and ITS by providing more accurate detection results to other autonomous vehicles and improve transportation safety.

C. Experiments on the ETH dataset

To extend our experiment, we directly apply the BCNet model and the baseline model CSP [25] trained on the CityPersons [4] dataset to test the ETH [9] dataset without any fine-tuning. The ETH dataset serves as an additional test dataset, and it has 1804 frames from three 15-FPS video clips. Fig. 6 is generated by the toolbox[1], which is the plotted miss rate against FPPI in log scale. In Fig. 6, our BCNet model yields slightly better results than CSP [25]. One reason why our model is not significantly better than the baseline model CSP [25] is because the resolution of the test image is only 640×480, which can weaken the effect of the visible part center score when generating the final heatmap. By comparing with other models trained or fine-tuned before evaluating on the ETH dataset, our model which only trained on the CityPersons dataset achieves a moderate performance on the ETH dataset, which proves the generalization of our model. With the advancement of the vehicular camera, the resolution and quality of images will be improved, which can benefit our model to locate center keypoints of pedestrians better. In this way, our model will show considerable performance even on the data it has never seen before.

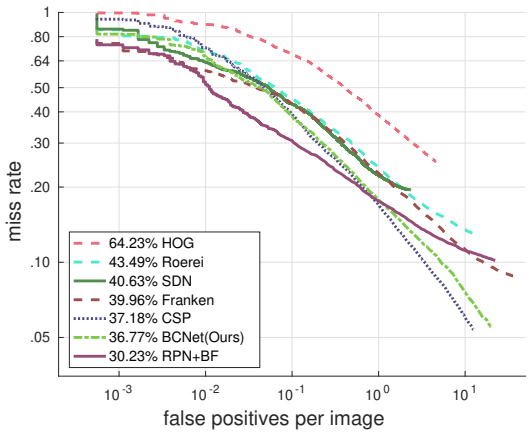


Fig. 6. Comparisons of the state-of-the-art models on the ETH dataset.

V. Conclusion

In this paper, we emphasized the importance of detecting pedestrians under reasonable occlusion with the support of vehicular networks and communications. We proposed BCNet that fuses the full body center keypoint prediction and the visible part center keypoint prediction for each pedestrian, thereby increasing the confidence score when the pedestrian is slightly occluded or unoccluded. We did the ablation study to find how different combinations of weights α and β would affect the performance. We tested our BCNet on the CityPersons dataset and the ETH dataset, and the results are promising, which will contribute to ITS by sharing more accurate detection results with other autonomous vehicles and transportation participants. In future work, we will improve our detector's accuracy on small resolution images, apply our detector to real-time video, and consider privacy protection.

[1]www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/index.html

References

[1] P. Dollar et al., "Pedestrian detection: A benchmark," in *Proc. IEEE CVPR*, 2009, pp. 304–311.
[2] H. Huang et al., "Widet: Wi-fi based device-free passive person detection with deep convolutional neural networks," in *Proc. ACM MSWiM*, 2018, pp. 53–60.
[3] P. Dollar et al., "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, 2012.
[4] S. Zhang et al., "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE CVPR*, 2017, pp. 4457–4465.
[5] X. Wang et al., "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE CVPR*, 2018, pp. 7774–7783.
[6] R. W. L. Coutinho et al., "Information-centric strategies for content delivery in intelligent vehicular networks," in *Proc. ACM MSWiM*, 2018, pp. 21–26.
[7] R. I. Meneguette et al., "A flow control policy based on the class of applications of the vehicular networks," in *Proc. ACM MobiWac*, 2017, pp. 137–144.
[8] L. Aliouat et al., "Flexible multipoint-to-multipoint routing protocol in ultra-dense nanonetworks," in *Proc. ACM MobiWac*, 2019, p. 81–87.
[9] A. Ess et al., "Depth and appearance for mobile scene analysis," in *Proc. IEEE ICCV*, 2007, pp. 1–8.
[10] D.G. Lowe et al., "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[11] N. Dalal et al., "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, pp. 886–893.
[12] C. P. Papageorgiou et al., "A general framework for object detection," in *Proc. IEEE ICCV*, 1998, pp. 555–562.
[13] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[14] R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE CVPR*, 2014, pp. 580–587.
[15] R. Girshick, "Fast r-cnn," in *Proc. IEEE CVPR*, 2015, pp. 1440–1448.
[16] S. Ren et al., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[17] J. Redmon et al., "You only look once: Unified, real-time object detection," in *Proc. IEEE CVPR*, 2016, pp. 779–788.
[18] W. Liu et al., "Ssd: Single shot multibox detector," in *Proc. ECCV*, 2016, pp. 21–37.
[19] T.-Y. Lin et al, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.
[20] S. Zhang et al., "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *Proc. ECCV*, 2018, pp. 637–653.
[21] C. Zhou et al., "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. ECCV*, 2018, pp. 135–151.
[22] T. Song et al., "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. ECCV*, 2018, pp. 536–551.
[23] H. Law et al., "Cornernet: Detecting objects as paired keypoints," in *Proc. ECCV*, 2018, pp. 734–750.
[24] K. Duan et al., "Centernet: Keypoint triplets for object detection," in *Proc. IEEE ICCV*, 2019, pp. 6569–6578.
[25] W. Liu et al., "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE CVPR*, 2019, pp. 5187–5196.
[26] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
[27] T. Song et al., "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. ECCV*, 2018, pp. 536–551.
[28] W. Liu et al., "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. ECCV*, 2018, pp. 618–634.
[29] S. Liu et al., "Adaptive nms: Refining pedestrian detection in a crowd," in *Proc. IEEE CVPR*, 2019, pp. 6459–6468.
[30] J.Deng et al., "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.
[31] D.P. Kingma et al. , "Adam: A method for stochastic optimization," [Online]. Available:https://arxiv.org/abs/1412.6980, 2014, accessed on: Sept., 2019.