Urban Scene Based Semantical Modulation for Pedestrian Detection

Hangzhi Jiang[a,b,*], Shengcai Liao[c], Jinpeng Li[c], Véronique Prinet[b], Shiming Xiang[b,a]

[a]*School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China*
[b]*Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[c]*Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE*

Abstract

Despite recent progress, pedestrian detection still suffers from the troublesome problems of small objects, occlusions, and numerous false positives. Intuitively, the rich context information available from urban scenes could help determine the presence and location of pedestrians. For example, roads and sidewalks are good cues for potential pedestrians, while detections on buildings and trees are often false positives. However, most existing pedestrian detectors ignore or inadequately utilize semantic context. In this paper, in order to make full use of the urban-scene semantics to facilitate pedestrian detection, we propose a new method called *Semantical Modulation based Pedestrian Detector (SMPD)*. First, for efficiency, a semantic prediction module is jointly learned with a baseline detector for semantic predictions. Second, a semantic integration module is designed to exploit the urban-scene semantic context for detection. Specifically, we force it to be an independent detection branch based solely on semantic information. In this way, together with the baseline detector, the fused detection results explicitly depend on both the learned appearance features and the scene context around pedestrians. In addition, while existing methods cannot be applied to the datasets where semantic annotations are not available for training, we introduce a semi-supervised transfer learning approach to make our method

*Corresponding author
Email address: jianghangzhi2018@ia.ac.cn (Hangzhi Jiang)

suitable for more scenarios. We demonstrate experimentally that, thanks to the integration of semantic context from urban scenes, SMPD can accurately detect small and occluded pedestrians, as well as effectively remove false positives. As a result, SMPD achieves the new state of the art on the Citypersons and Caltech datasets.

Keywords: Pedestrian Detection, Semantic Context, Urban Scene

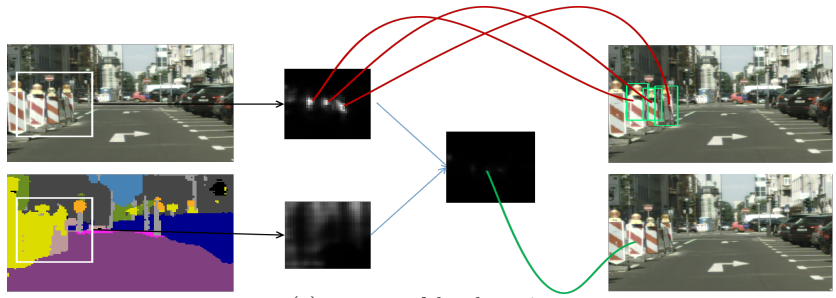
1. Introduction

Pedestrian detection is a key component in intelligent video analysis and has wide applications in surveillance systems, intelligent transportation, driving assistance, etc. However, it is still a difficult task in complex scenarios, with many challenges such as small objects, occlusions of various levels, numerous false positives, etc.

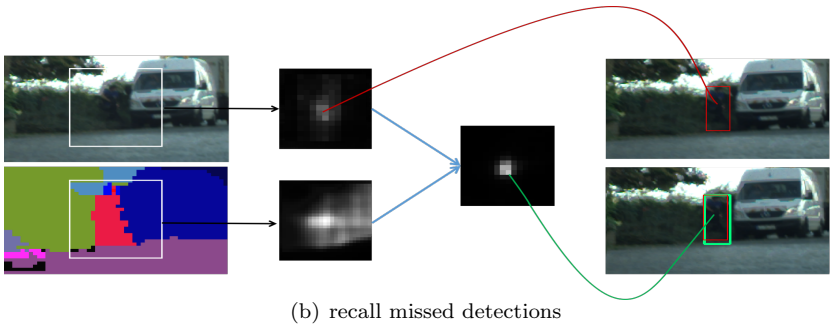
In recent years, convolutional neural network (CNN) models [1, 2, 3, 4, 5] have made great progress in generic object detection. Along this line, various pedestrian detectors [6, 7, 8, 9, 10] have been designed based on CNN and demonstrated improved performance. However, most of these methods localize objects merely using learned appearance features, which may be insufficient for addressing the above challenges.

Considering the value of context, several methods [8, 11, 12] have been proposed to incorporate features of larger regions as context for object detection. However, the context used in these methods lacks clear semantic meaning, thus also insufficient to facilitate detection. Lately, motivated by the close connection between detection and segmentation, various works [13, 14, 15, 16, 17] focus on improving the detection accuracy using auxiliary object semantics. However, these methods only exploited semantic pixels of various foreground instance objects while regarded background as a single class, ignoring the rich scene information in the background that can be used as detection cues.

In contrast to the foreground semantics, the urban-scene semantics around pedestrians do provide rich context information, which could be helpful for



(a) suppress false detections



(b) recall missed detections

Figure 1: Two examples of how scene semantics predicted in SMPD can help detection. Red and green boxes represent the groundtruth and detection results, respectively. The images at the four corners represent original images, detections of the baseline detector, semantic maps, and detections of SMPD, respectively, from top left to bottom right. The middle of the two subgraphs shows probability maps for detection, where the maps in the first column are from the baseline detector and the semantic maps, respectively, and that in the second column are the semantical modulation results by multiplying these two probability maps.

pedestrian detection if utilized properly. For example, streets and sidewalks are good cues for potential pedestrians, while detections on the sky and trees are typically false positives. With this in mind, Zhang *et al.* [7] proposed to facilitate detection by introducing an additional segmentation map containing scene semantics as the fourth channel of the image. However, the segmentation maps were produced by an independent network, requiring heavy additional computation. Mao *et al.* [18] and Chen *et al.* [19] further proposed to jointly learn semantic segmentation and detection in a single network. However, in these two methods, regardless of the predicted semantics, the CNN features

of the semantic branch were used and simply concatenated with features of the backbone to improve detection. This feature-level fusion processing makes it ambiguous that how much useful semantic information is contained in the features used for final detection. In other words, it is unable to intuitively and fully utilize semantic information for detection. Actually, the complementary detection clues contained in the scene semantics are sufficient, so that networks can explicitly and directly use the semantic context to guide detection. And we do show in this work that implicitly using CNN features from the semantic branch and the simple feature concatenation do not fully exploit the rich urban-scene context, while explicitly utilizing the semantic class probability predictions as complementary detection cues to directly guide detection is a better choice.

Therefore, motivated by the rich detection cues contained in the background scene, and for addressing the problem that most existing pedestrian detectors ignore or inadequately utilize semantic context, we propose a novel method to facilitate pedestrian detection through an intuitive and effective use of scene semantics with explicit semantic meanings, denoted as *Semantical Modulation based Pedestrian Detector (SMPD)*. Specifically, in SMPD, the semantic predictions are independently used to perform pedestrian detection through a *semantic integration module*. The final detection results are determined by fusing the outputs of semantic integration module and the baseline detector, which forms a dual detection structure. We call this procedure *semantical modulation*. For efficiency, the semantic predictions are generated by a *semantic prediction module*, which is jointly learned with the detection network. The two modules are laterally included into an existing detector, requiring few additional parameters and little computational cost. Through the dual detection structure, the detection results explicitly depend on both the learned appearance features and the scenes in which the objects are located. It fully utilizes the constraint and complementary functions of scene semantics. In addition, semantic labels are not available for most pedestrian detection datasets and existing pedestrian detectors using semantic information cannot be effectively applied to the datasets without semantic labels. Therefore, considering the low requirement of

the proposed network for the segmentation precision, a semi-supervised transfer learning approach for generating pseudo semantic labels is further introduced. Experimentally, our method can be successfully trained with the pseudo semantic labels generated by this approach, without the requirement of additional semantic annotations. It makes our method suitable for more scenarios and distinct from existing methods always requiring semantic annotations.

Thanks to the effective use of urban-scene semantics, SMPD achieves high performance in detecting small and occluded pedestrians, as well as removing false positives. As can be seen from the two examples in Fig. 1, the baseline detector introduces false positives and fails to detect a pedestrian. In contrast, when integrating the detection results obtained from the predicted semantic maps, the probabilities of the false positives and the missed pedestrian are suppressed and enhanced, respectively. Thus, pedestrians are successfully detected by SMPD. In a quantitative evaluation, the proposed SMPD achieves new state of the art on the Citypersons [7] and the Caltech [20] datasets.

2. Related Work

Object detection: With recent development of CNNs, generic object detection has gained great success. In general, various CNN-based detectors can be roughly divided into multi-stage, single-stage and anchor-free methods. The multi-stage methods [1, 3, 21, 22, 23] first generate object proposals, and then refine these proposals using sub-networks for classification and regression. Inspired by these, numerous methods have tried to improve detection performance by focusing on the network architecture [21, 22, 24], training strategy [23, 25], etc. The single-stage methods [4, 5, 26] aim at speeding up the detection process by directly classifying and regressing anchor boxes. Though faster, the detection accuracy of single-stage detectors often lags behind the two-stage methods. To improve the detection accuracy, Dssd [27] focused on enhancing the feature representation, while RetinaNet [28] paid attention to the extreme positive-negative imbalance problem during training. Recently, anchor-free methods [29, 30, 31]

attract more attention due to their simple structures without parameter settings of anchors and achieve high performances. For example, CornerNet [29] proposed to classify the upper left and lower right corners of objects in a pixel-wise manner and match the corners through embedding. FCOS [31] proposed to locate the sample points in the center area of each object and regress the distances from the sample to the boundaries, thus replacing the pre-defined anchors. However, when applied directly to the field of pedestrian detection, all above methods have not presented competitive performances, especially on small or heavily-occluded pedestrians.

Pedestrian detection: Traditional pedestrian detectors, such as ACF [32] and LDCF [33], exploited various filters on Integral Channel Features [34] to localize objects. Recently, inspired by the success of CNNs in generic object detection, various methods [35, 9, 6, 36] have been designed for pedestrian detection. For example, DeepParts [37] first applied traditional method [33] to generate proposals, and then employed a CNN for classification. In contrast, NeuralFeatures [38] combined neural features from FCN [39] with a traditional AdaBoost classifier. RPN+BF [6] applied a stand-alone RPN for proposal generation, and a boosted decision forest for classification. And HCD [40] further extended RPN+BF framework to combine handcrafted features and CNN features for detection. To address the multi scale problem, SA-FastRCNN [9] jointly trained two networks to detect large and small pedestrians, respectively. MS-CNN [8] exploited different layers to generate proposals corresponding to different scale ranges. MHN [41] proposed a multi-branch network to generate multiple similarly high-level feature maps of different resolutions. AMS [42] utilized asymmetric rectangular convolution kernels for capturing the compact features of pedestrians. SML [43] enhanced the features of small objects through forcing the feature representations of small-scale pedestrians to approach those of large-scale pedestrians. For the occlusion problem, RepLoss [10] designed a regression loss to make predictions belonging to different groundtruths far away from each other. HBAN [44] conducted head detection in parallel with traditional body branch. Zhang *et al.* [45] further introduced body part detec-

tion. Adaptive-NMS [46] assigned different Non-Maximum Suppression (NMS) thresholds for objects according to densities. PBM [47] further utilized the predicted visible boxes in the process of NMS. Considering the speed, ALFNet [48] proposed to directly regress anchor boxes in a multi-step process. CSP [49] was further proposed as an anchor-free detector, which only predicted the center and scale maps, achieving state-of-art results on Citypersons [7]. Due to the advantages of CSP in both speed and accuracy, we utilize it as the baseline detector in our experiments

Object detection with semantic segmentation: Due to the complementary information in object semantics, various works [50, 15, 17, 14, 51, 13, 52] focus on the effectiveness use of the object semantics for detection. F-DNN [50] segmented pedestrians through an independent network to refine the detection results in a post manner. Brazil *et al.* [15] exploited foreground and background information, which was combined with the features from the backbone. SSA-CNN [53] explored pedestrian segmentation results as self-attention cues to boost pedestrian detection. MGAN [17] and MDFL [16] further proposed to segment the coarse visible parts of objects as foreground for filtering features through attention. Mask R-CNN [13] achieved instance segmentation from shared feature maps on top of Faster R-CNN. However, these methods regard the background as a single class, ignoring the rich scene information that could be used as detection cues. There are also various works [54, 55, 7, 18] that do use scene semantics for pedestrian detection. For example, Zhang *et al.* [7] and Sheng *et al.* [55] improved detection by combining the segmentation maps with the original images and traditional hand-crafted features, respectively. However, the segmentation maps were predicted by an independent network [39], thus introducing extra computational burdens. Mao *et al.* [18] jointly learned segmentation and detection, combining features learned by segmentation with feature maps in higher layers. However, the simple feature concatenation in these methods prevented the semantic context from being fully exploited. Different from these works, SMPD pays more attention to how the urban-scene semantics can be used for pedestrian detection, which performs individual detections based on

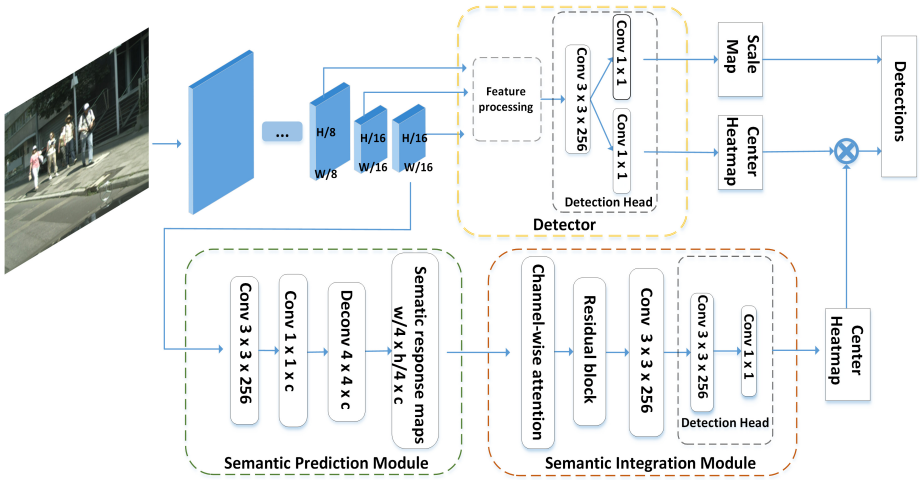


Figure 2: The overall architecture of SMPD is based on a baseline detector that consists of a feature processing module and a detection head. Apart from the baseline detector, SMPD consists of two main components, i.e. the Semantic Prediction Module (SPM) and the Semantic Integration Module (SIM). The SPM generates semantic response maps of size $w/4 \times h/4 \times c$, where c is the number of semantic categories. The SIM produces a center heatmap by feeding detection features exploited from semantic response maps to the detection head. The final detection results are determined by the outputs of the two detection heads.

the semantic information exploited from urban-scene semantic maps for constraining and supplementing the detection results. In addition, existing works all require semantic annotations for training. In contrast, we present a solution in Section 3.4 when no semantic labels are available for training.

3. Approach

3.1. Overall Architecture

The overall architecture of the proposed SMPD is illustrated in Figure 2. ResNet-50 [56] is chosen as the backbone for extracting features. Since the resolution of feature maps in the final stage of the backbone are too low for small objects, as is common practice, dilated convolutions are adopted in the last stage to make the feature maps downsample to 1/16 the size of the input

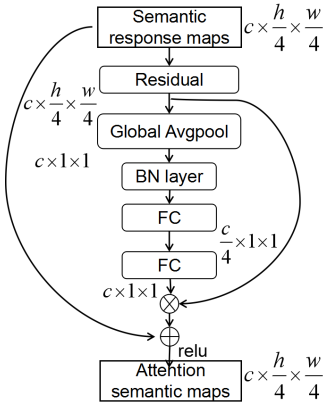


Figure 3: The structure of the SE+BN module used in SIM.

images. Considering accuracy and speed, SMPD uses CSP[48] as the baseline pedestrian detector, which is an anchor-free detection framework. CSP first fuses multi-scale feature maps, and for object detection, a detection head is appended on the concatenated features to produce the center heatmap and scale map. For a fair experimental comparison and better performance, SMPD selects the feature maps from stage 3-5 in ResNet-50 as the multi-scale feature maps, which is consistent with the best performing model of CSP.

As depicted in Figure 2, apart from the baseline detector, SMPD has two key modules, namely *Semantic Prediction Module (SPM)* for generating semantic response maps from the shared feature maps, and *Semantic Integration Module (SIM)* for independent detections based on the semantic context. The SPM is connected to the final outputs of the backbone while the SIM is applied on the semantic response maps. The final classification results are determined by the product of the classification maps from the SIM and the basic detector, and we call this procedure *semantical modulation*. In the following sections, we will present the details and design principles of the SPM and the SIM. And for showing the structure of the proposed modules more clearly, the detailed parameters of the proposed SPM and SIM are listed in Table 1 of the supplementary materials.

3.2. Semantic Prediction Module (SPM)

The semantic prediction module is utilized to produce pixel-level semantics, which can provide useful information to facilitate pedestrian detection. Generally speaking, high-level feature maps have more semantic information. Therefore, SPM is appended upon the final feature maps of the backbone, which are 1/16 the size of the input images.

The structure of the SPM is depicted in Figure 2, where two buffering convolutional layers are first used to prevent the gradients of the SPM branch from being back-propagated directly to the backbone and causing instability during joint training. After the buffering layers, the number of feature channels is reduced to c, where c is the number of semantic categories, which is not usually very large. Second, for further detection with requirements on small objects, semantic response maps are enlarged to 1/4 size of the input images via a deconvolutional layer. As the outputs of SPM, the semantic response maps are supervised by the semantic labels during training. In contrast to other detectors that only use two categories (foreground and background) for facilitating detection, SMPD utilizes various semantic scene categories (20 classes in Cityscapes [57]), and can thus fully exploit the rich context. Finally, the semantic response maps are fed into the SIM.

3.3. Semantic Integration Module (SIM)

The semantic integration module aims at fully exploiting and using the scene information from the semantic response maps for pedestrian detection. Intuitively, scene semantics contain rich detection clues, the semantic predictions can be directly used to make decisions on the existence probability of objects. Motivated by this, different from the implicit feature fusion scheme in previous works [19, 18], the SIM branch individually estimates pedestrian bounding box candidates directly from the feature maps exploited from the semantic response maps with explicit semantic meanings. Figure 2 shows the structure of the SIM.

In the design of SIM, a channel-wise attention is first appended upon the semantic response maps. It has a structure similar with the existing Squeeze-and-

Excitation (SE) module [58] and assigns different weights for different channels. This is done for two reasons. On the one hand, in order to make correct decisions of detection based on the semantic context, SIM should learn the relationship between the semantics of different categories and the pedestrian semantic. Taking the 20 categories in Cityscapes as an example, in addition to the pedestrian, the road and sidewalk are scenes positively related with pedestrians, while the sky and traffic signs are negatively correlated. In contrast, the presence of wall or fence is not correlated with pedestrians. On the other hand, if using deeper stacks of convolutional layers instead, the difference in importance of different categories may also be implicitly learned, but this would increase the computational cost. Therefore, a SE module for channel-wise attention should be useful to endue different weights to different semantic categories and reduce the learning difficulty, as well as the network complexity. However, since the input of SIM contains explicit semantic meaning rather than unconstrained features, and the total pixels of different semantic categories are usually imbalanced. And due to the global average pooling in the SE module, the channels with larger response pixels will dominate the weight calculation. Therefore, the learning difficulty will be significantly increased if using the original SE module. In addition, only 1/16 channels are left after the first FC layers in the original SE module, which is harmful to the relation extraction between few semantic categories. Thus, as shown in Figure 3, after the first FC layer, 1/4 channels are reserved, which allows the SE module to better extract relations between different semantic categories. Besides, BN layer is introduced after the Global Pooling to handle the trouble of imbalanced semantic categories. Because this operation can ensure that the pooled values of different categories are normalized to the same range. We denote this adapted module 'SE+BN'.

For further detection, after 'SE+BN', we attach a residual block to the adjusted semantic maps to increase its dimension to 256 and then append a 3x3 conv layer to generate the semantic-based features for detection. The obtained semantic-based feature maps are finally fed into the detection head to produce the center heatmap. In this way, the detection heads in the baseline detector

generate predictions using only the appearance features, while the SIM generates detection scores according to the rich urban-scene semantics surrounding the predicting points. And the final detection results are determined by the product of the two center heatmaps. Therefore, this semantical modulation detection approach makes it easier for the network to detect pedestrians, especially in ambiguous cases like small or occluded objects, or person-like blobs in unreasonable places. When training, the product of probability maps from the two detection heads is supervised by the classification labels.

3.4. Training

Loss functions. SMPD is optimized in a joint training manner with two main objectives for semantic prediction and detection. For semantic prediction, the loss function l_s is formulated as:

$$l_s = \sum_{i,j} \sum_{c=1}^{C} -w_c \hat{Y}_{ij}^c \log Y_{ij}^c, \tag{1}$$

where $\hat{Y}_{ij}^c$ and Y_{ij}^c are the groundtruth and the predicted semantic probability, respectively, for the cth category at location (i, j). Y_{ij}^c is obtained by applying the softmax function on the outputs of SPM along the category channels. Following [59], w_c is the weight of the cth category to balance the segmentation loss contribution of different categories, which is inversely proportional to the ratio of positives in the cth category to the total number of pixels. Note that the original semantic labels should be downsampled to the size of the semantic response maps that is supervised by the l_s.

For object detection, following most methods [7, 48, 49], the classification loss l_{cls} and the regression loss l_{loc} are applied. The focal loss [28] used for l_{cls} is defined as follows:

$$l_{cls} = -\alpha \sum_{(i,j)\in S_+} (1 - p_{ij})^\gamma \log(p_{ij}) - (1 - \alpha) \sum_{(i,j)\in S_-} {p_{ij}}^\gamma \log(1 - p_{ij}) \tag{2}$$

where p_{ij} is the classification probability of pedestrian center at location (i, j), S_+ is the classification positive set containing all the points where the centers of

pedestrians are located, S_- is the negative set containing rest points, and α and γ are focusing parameters, set to 0.25 and 2, respectively, as suggested in [28]. In this way, the loss contributions of easy samples are down-weighted. During training, l_{cls} is applied to supervise the product of the two center heatmaps from the SIM and baseline detector.

Generally, the pedestrian bounding box can be regarded as having a fixed aspect ratio. And in the baseline detector CSP, positive samples for classification are the object centers. Therefore, CSP directly regresses the height of the object. Following CSP, the proposed SMPD also regresses the height and sets the aspect ratio as 2.4. For supervising the height predictions, we adopt the smooth L1 loss as the regression loss l_{loc}:

$$l_{loc} = \frac{1}{K}\sum_{k=1}^{K} SmoothL1(h_k, t_k), \tag{3}$$

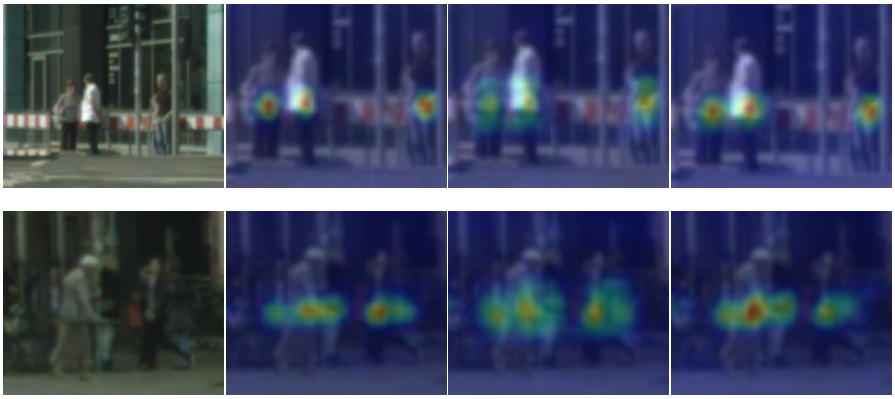
where h_k and t_k represents the predicted and groundtruth height at the center point of the kth bounding box, respectively. Note that the Equation 3 is only used to supervise the height predictions of positives.

To sum up, the total loss function is formulated as:

$$L_{total} = \lambda_c l_{cls} + \lambda_l l_{loc} + \lambda_s l_s, \tag{4}$$

where $\lambda_c, \lambda_l, \lambda_s$ are the weights of the classification loss, regression loss and semantic prediction loss, respectively. Following CSP, λ_c, λ_l are set as 0.02 and 1. λ_s is set as 0.1. These weights keep the three losses on the same magnitude.

Gaussian map. In general, all pixels with semantic pedestrian labels are positives for segmentation, while for CSP, only the pedestrian centers are positives for detection. Thus when the segmentation loss in Equation 1 is appended, all positive pedestrian pixels around object centers are treated equally for segmentation. This will introduce more difficulties for the network to learn the salient features of center. This feature interference will reduce the benefits of multi-task learning for the center-based detectors. As can be seen in the Figure 4(b) and 4(c), the features of the baseline detector can clearly highlight the object center. However, when only introducing SPM, the features of the object



(a) input image (b) baseline (c) baseline+SPM (d)baseline+SPM+GM

Figure 4: Two examples of features of different models before the final 1x1 conv layer in the detection head. The images from (b) - (d) represent the features of the baseline, baseline+SPM and baseline+SPM+GM, respectively. The features visualized here are obtained by norming the feature map along the channels, in which brighter means larger response.

center become fuzzy and have low response. To address the feature interference, we design a Gaussian Map (GM) that enables to gradually decrease segmentation responses for pedestrian pixels located away from centers. The semantic pedestrian prediction Y_{ij}^p in Equation 1, where p represents the pedestrian category, can be weighted as $\widetilde{Y}_{ij}^p = g_{ij} Y_{ij}^p$, where g_{ij} is the weight at location (i, j) in the Gaussian Map. It is formulated as follows:

$$g_{ij} = \begin{cases} 2 - \max\limits_{k} G(i, j, x_k, y_k, \sigma_{w_k}, \sigma_{h_k}), & (i, j) \in S, \\ 1, & \text{otherwise}, \end{cases} \tag{5}$$

where (x_k, y_k) are the center coordinates of the kth bounding box, S is the union area of all the bounding boxes, $(\sigma_{w_k}, \sigma_{h_k})$ are the variances of the Gaussian distribution, which are proportional to the height and width of individual objects. $G(i, j, x_k, y_k, \sigma_{w_k}, \sigma_{h_k})$ is the density value at location (i, j) in the Gaussian distribution centered at (x_k, y_k) with variances $(\sigma_{w_k}, \sigma_{h_k})$, which adopts the same

setting as the Gaussian mask in [49] and can be formulated as:

$$G(i,j,x,y,\sigma_w,\sigma_h) = e^{-(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(j-y)^2}{2\sigma_h^2})}, \tag{6}$$

Finally, Y_{ij}^p in Eq. 1 will be replaced by $\widetilde{Y}_{ij}^p$ for training. Note that this is done only in the pedestrian channel, but not in channels of other categories. In this way, larger weights are given to pedestrian pixels further away from their centers for segmentation, resulting in a lower response of these points, and thus avoiding interference with the center detection in CSP. As shown in Figure 4(d), introducing GM can be competent to alleviate the feature interference. It makes the features used for detection brighter and more concentrated than that of the baseline+SPM, while retaining the semantic information.

Semi-supervised transfer learning. Existing methods that use semantic information for pedestrian detection can not be applied to datasets where semantic annotations are not available. To tackle this issue, we introduce a semi-supervised transfer learning approach which generates pseudo semantic labels for training. Specifically, in this method, the detection labels are also used to help revise the pseudo semantic labels of the pedestrian channel. The process of the semi-supervised transfer learning approach to generate pseudo semantic labels can be described by the following four steps:

(1) For an image, an additional pre-trained segmentation model firstly generates semantic predictions with the size of $W \times H \times C$, where C is the number of semantic categories.

(2) To make the pedestrian segmentation more accurate, we generate a foreground and background mask ($W \times H \times 1$). In this mask, all the pixels within the pedestrian boxes are regarded as foreground and set to the same value (larger than 1), and all other pixels are set to 0.

(3) The pedestrian channel of the semantic predictions is multiplied by the mask to produce adjusted semantic predictions.

(4) Finally, the pseudo semantic labels are obtained through arg max on the channel dimension of the adjusted semantic predictions.

In Section 4.4, we show the success of this learning approach. Theoretically, the success of this method is inseparable from the low requirement of the proposed SMPD for segmentation precision. There are two reasons for the low requirement. Firstly, the baseline detector and SIM perform detection separately and only interact on the final center heatmaps. In this way, the detection results of the baseline detection head can still mainly depend on the learned appearance features. This reduces the interference caused by the low-precision segmentation map to the features used for detection of the baseline detector. Secondly, SIM is a detection branch, and the detection task mainly needs salient features which integrate local context information. And due to the use of multi-layer 3x3 convolution in SIM, the receptive field for detection is expanded. Therefore, SIM relies on the recapitulative semantic context to preform detection, rather than the semantic of a single pixel.

	Caltech		Citypersons	
	Train	Val.	Train	Val.
# images	42782	4024	2975	500
# persons	13674	1358	19654	3938
# ignore regions	50363	6238	6768	1631
# person/image	0.32	0.34	6.61	7.88

Table 1: Statistics on CityPersons and Caltech dataset.

4. Experiments

4.1. Datasets

The performance of SMPD is evaluated on Citypersons [7] and Caltech [20] datasets. The detailed statistics on the two dataset are listed in Table 1.

Citypersons. Citypersons is annotated on the fine annotation images in Cityscapes [57] dataset. The dataset contains 2975 images and approximately 20000 pedestrians in the training subset. The proposed model is trained on this

subset, with corresponding semantic labels from Cityscapes and evaluated on the validation subset with 500 images. Cityscapes is a segmentation dataset of urban scenes, which is recorded across 50 different cities in Germany with 3 different seasons and various weather conditions. The dataset contains 5000 images with fine annotations and defines 19 semantic categories for evaluation. In training, we choose these 19 semantic categories and an additional background class for the loss calculation.

Caltech. The Caltech pedestrian dataset consists of approximately 10 hours of 640x480 30Hz video taken from a vehicle driving in an urban environment. The dataset contains 11 sets of videos, in which the sets 0-5 are used for training and the sets 6-10 are for testing. Following [49, 17, 48, 46, 7, 18], we use the 10x set (42782 images) as the train subset which samples the frames at 10Hz, and the test subset (4024 images) for evaluation. Our SMPD is trained and tested with the new annotations [60] on the original image size.

Evaluation Metric. The evaluation follows the standard Caltech evaluation metric [20], that is the log-average Miss Rate over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}). The lower value of MR^{-2} reflects better detection performance. Results on ignored regions will not be considered in the evaluation.

4.2. Training details

We choose the ResNet-50 pretrained on ImageNet [61] for backbone. All other layers are randomly initialized with the 'Xavier' method. For fair comparison, SMPD applies the same positive-negative definition as CSP [49]. In addition, following CSP, the training strategy of moving average weights proposed in [62] are also applied to achieve more stable training. To increase the diversity of the training data, the standard data augmentation techniques including random color distortion, random horizontal flip, random scaling and random crop are adopted. Since there are no semantic annotations for Caltech, the pseudo labels generated by the semi-supervised transfer learning approach are used for training, as described in Section 3.4. In the transfer learning approach, the

Method	SPM	SIM			GM	MR^{-2}(%)		Parameters	Speed
		w/o SE+BN	w SE	w SE+BN		IoU=0.5	IoU=0.75	(MB)	(s/img)
Baseline						11.02	41.76	38.1	0.15
SMPD	✓					11.30	37.86	42.6	0.16
	✓				✓	10.69	36.82	42.6	0.16
	✓	✓				10.64	37.51	44.3	0.17
	✓		✓			11.06	39.31	44.3	0.17
	✓			✓		10.01	35.99	44.4	0.17
	✓	✓			✓	10.28	36.37	44.3	0.17
	✓			✓	✓	**9.89**	**34.58**	44.4	0.17
gain						+1.13	+7.18		

Table 2: Contribution of each component, evaluated on Citypersons [7]. In the SIM column, w/o SE+BN, w SE and w SE+BN represent the SIM equipped without SE module, with original SE module and with the 'SE+BN', respectively.

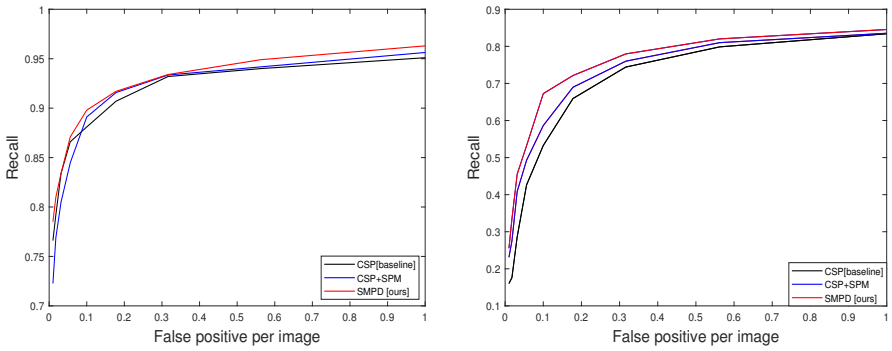
segmentation model [63] used for producing the semantic predictions is trained on Cityscapes. In the following experiments, unless otherwise stated, models are trained and tested on the original image size. For more details regarding the configurations on Citypersons and Caltech, please refer to [49].

We implement SMPD on Pytorch platform. For Citypersons, the network is trained for 40k iterations by the Adam optimizer on two Tesla V100 GPUs with 8 images per GPU. The initial learning rate is 2×10^{-4} and decreases to 2×10^{-5} after 30k iterations. For Caltech, the network is trained for 10k iterations on one GPU with a batchsize of 16 by the Adam optimizer. The learning rate is set to 2×10^{-4}. Following [48, 49, 45], we also conduct experiments on Caltech with the model pretrained on Citypersons. And the model is trained for 5k iterations with a learning rate of 10^{-4}.

4.3. Ablation Study

In this section, we report ablation studies on the Citypersons dataset [7].

Component evaluation. To evaluate the effectiveness of each component of SMPD, we train the models with different components starting from the baseline CSP. As shown in Table 2, CSP achieves 11.0% MR^{-2} under IoU=0.5,



(a) IoU=0.5 (b) IoU=0.75

Figure 5: ROC curves of the baseline CSP (black), CSP with SPM (blue), and SMPD (red) under IoU = 0.5 and 0.75, evaluated on Citypersons.

which is still a leading result on Citypersons under the original image size. Based on it, jointly training SPM results in a small performance decrease at IoU=0.5, but an improvement at IoU=0.75, These results are consistent with the ROC curves in Figure 5. Empirically, the evaluation metric under IOU=0.5 is more concerned with the ability of the model to distinguish between positives and negatives. Because in this case, if one positive point is classified correctly, it will be regarded as a true positive as long as the IOU between its box prediction and corresponding gt is greater than 0.5. In contrast, the evaluation metric under IOU=0.75 emphasizes the regression quality of the positive samples. Therefore, the most intuitive explanation of these results is that jointly training baseline with SPM leads to a better regression ability, but worse classification ability. As discussed in Section 3.4, the positive pedestrian pixels for segmentation degrade the shared feature being discriminative for center localization, but the scale regression is more accurate due to the boundary information involved from the semantic predictions. To address the feature interference in learning, adding Gaussian Map enables to reduce the MR^{-2} to 10.69%. When only the SIM is added without 'SE+BN', the MR^{-2} is reduced to 10.64%, which indicates the effectiveness of the dual detection scheme in using the scene semantics. However,

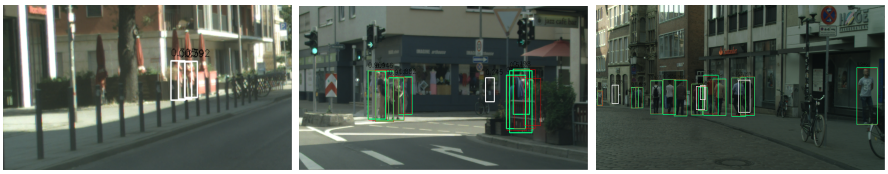
Method	[50,75]	[75,100]	[100,∞]
Baseline	16.0	3.7	6.5
SMPD	**14.8**	**2.6**	**5.6**
improvement	+1.2	+1.1	+0.9

Table 3: Comparison on MR^{-2} between the baseline and SMPD in different object scales under IoU = 0.5 on Citypersons.

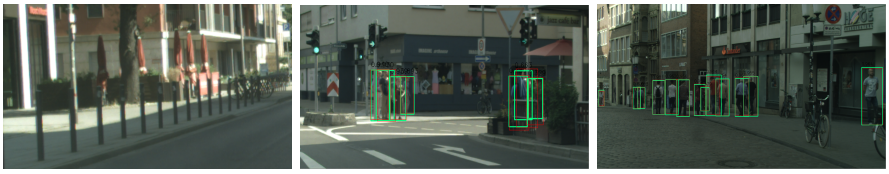
Method	Categories	Integration	IoU=0.5	IoU=0.75
Baseline	-	-	11.02	41.76
SMPD	-	dual detection	10.93	38.14
	2	dual detection	10.70	39.32
	20	feature concat	10.71	38.68
	20	dual detection	**10.28**	**36.37**

Table 4: MR^{-2} performance on Citypersons of SMPD with different numbers of semantic categories and different methods in integrating scene semantics. For a fair comparison, all the SMPD models here are without 'SE+BN' module, except the model without segmentation supervision in the second line.

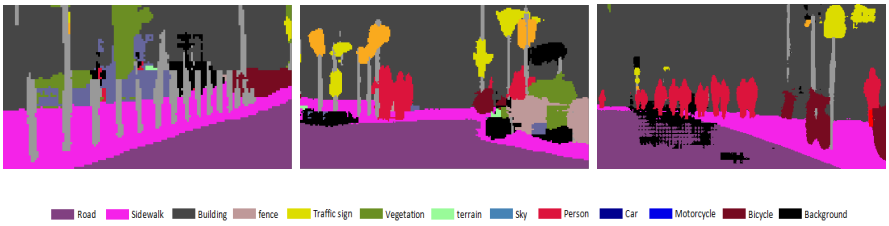
when the original SE module is used in SIM, the performance degenerates due to the imbalanced total number of pixels in different categories. And when 'SE+BN' is further enabled, the MR^{-2} is reduced to 10.01%, which indicates that, upon the semantic predictions, the proper use of the SE module with BN layers can reduce the difficulty of the SIM learning and exploit more effective features from the scene semantics for detection. Finally, all the effective modules are appended together, which achieves 9.89% MR^{-2} with a 1.13% improvement than the baseline. Under a stricter IoU of 0.75, SMPD achieves an even larger performance gain of 7.18% MR^{-2}, which indicates that it is capable of achieving a better localization quality with little additional computational cost.



(a) detection results of CSP



(b) detection results of SMPD



(c) prediction results of semantic segmentation from SMPD

Figure 6: Detection examples of the baseline CSP and SMPD on Citypersons. Red, green and white rectangles represent groundtruth, true positives and false positives, respectively. (a) detection results of CSP, (b) detection results of SMPD, (c) segmentation results produced by SMPD.

Different object scales. To demonstrate the performance on various object scales, we evaluate SMPD in three object scales according to [7]. As shown in Table 3, SMPD improves the results on all three scales. Specifically, the improvements on small ([50,75]) objects and medium ([75,100]) objects are encouraging, with MR^{-2} reduced by 1.2% and 1.1%, respectively, which indicates that the context information exploited from urban-scene semantics is quite beneficial to detect smaller objects.

False positive rate. SMPD aims at fully exploiting the semantic dependencies between pedestrians and the surrounding scenes, which can indicate whether there is a high probability of pedestrians. To verify this view, an experiment is conducted to demonstrate the ability of our SMPD model to reduce false positives. As shown in Figure 5, under both IoU=0.5 and 0.75, SMPD performs consistently better than CSP. It means that SMPD can achieve a clearly lower false positive rate.

To further demonstrate the effectiveness of SMPD in reducing false positives, we visually illustrate detection examples of CSP and SMPD in Figure 6. It can be observed that SMPD has fewer false positives. In the second and third images, SMPD does not generate false positives on buildings or between crowds, thanks to the contextual semantics in the corresponding semantic maps shown in Figure 6(c). Moreover, SMPD also reduces false positives with similar shapes to pedestrians such as the poles in the first image. It can also be observed that SMPD obtains better performance on occluded objects, such as the detected pedestrian on the right of the second image by means of the semantic context of the occlusion scenes.

Multi-category scenes. To verify how much the improvement in detection performance comes from the effectiveness of multi-category scene semantics, we compare the method with foreground-only and multi-category semantics. Foreground-only means that only the pedestrian and background categories are utilized in the SPM. As shown in Table 4, compared with CSP, SMPD with foreground-only semantics has a small performance gain under IoU=0.5. In contrast, with multi-category semantics, the MR^{-2} is largely improved. These results indicate that various semantic categories not only provide fore-background constraints but also enable the network to learn the relations between pedestrians and the semantic scenes around them. In order to verify whether the performance gain comes from the effective use of the scene semantics by SMPD or from the model with more parameters, we add an experiment. In this experiment, all the modules of SMPD are kept, but during training, there are no segmentation labels to supervise the outputs of SPM. As shown in the second

line in Table 4, the result of this model is only 10.93% MR^{-2}. The result stems from the fact that, without the supervision of semantic labels, the combination of SPM and SIM can not exploit supplementary clues for the baseline detector. This result illustrates the importance of various semantic categories to the proposed SMPD again. From another aspect, it also confirms that the structure design of SMPD and the multi-category scene semantics complement each other.

Detection branch based on semantic predictions. The major performance gain brought by SMPD lies in using the semantic predictions as detection cues for additional detection. And compared with the simple feature concatenation commonly used in other works, this dual detection structure can exploit semantic context more sufficient for detection. To demonstrate this, an additional comparison is shown in Table 4, where the features exploited in the SPM are concatenated with the features learned in the baseline detector, instead of the dual detection scheme. As can be seen, the MR^{-2} of the simple feature concatenation is only 10.71%, indicating that the dual detection scheme in SMPD is more effective.

4.4. Comparison with the State of the Art

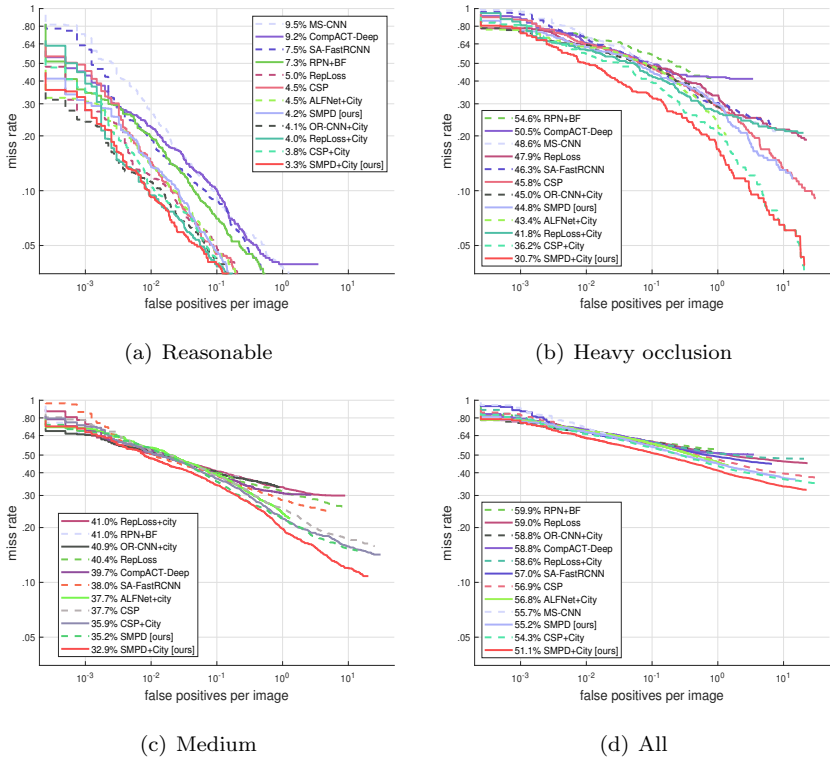
Citypersons. Table 5 shows a comparison with the state of the art on Citypersons [7]. In addition to the reasonable subset, SMPD is also evaluated on three subsets with different occlusion levels. It is worth noting that there are two division standards for the heavy occlusion subset. The visible range of pedestrians corresponding to the two standards is [0, 0.65] and [0.2, 0.65], respectively. On the reasonable subset, SMPD achieves the best performance, with an improvement of 0.6% MR^{-2} compared with the closest competitor MGAN, under the $\times 1$ image scale. This is even better than all methods tested on the $\times 1.3$ image size. On various occlusion levels, SMPD also performs quite well. Specifically, under the two standards for heavy occlusion and the partial occlusion subset, SMPD achieves 45.6%, 36.6% and 9.0% MR^{-2}, even better than RepLoss and OR-CNN which are specifically designed for the occlusion cases. These results demonstrate the ability of SMPD in handling occlusion issues, due

Method	Scale	Backbone	Reasonable	Heavy		Partial	Bare
				vis=[0,0.65]	vis=[0.2,0.65]		
Adapted Faster R-CNN [7]	×1	VGG16	15.4	-	-	-	-
Adapted Faster R-CNN+Seg [7]	×1	VGG16	14.8	-	-	-	-
Adapted Faster R-CNN [7]	×1.3	VGG16	12.8	-	-	-	-
TLL [64]	×1	ResNet-50	15.5	53.6	-	17.2	10.0
TLL+MRF [64]	×1	ResNet-50	14.4	52.0	-	15.9	9.2
Repulsion Loss [10]	×1	ResNet-50	13.2	56.9	-	16.8	7.6
	×1.3	ResNet-50	11.6	55.3	-	14.8	7.0
Bi-box [65]	×1.3	VGG16	11.2	-	44.2	-	-
OR-CNN [45]	×1	VGG16	12.8	55.7	-	15.3	6.7
	×1.3	VGG16	11.0	51.3	-	13.7	**5.9**
ALFNet [48]	×1	ResNet-50	12.0	51.9	-	11.4	8.4
PBM [47]	×1	VGG16	11.1	-	53.3	-	-
FRCN+A+DT [66]	×1.3	VGG16	11.1	-	44.3	11.2	6.9
Adaptive-NMS [46]	×1.3	VGG16	10.8	54.0	-	11.4	**6.2**
CrowdDetect [67]	×1.3	ResNet-50	10.7	-	-	-	-
CSP [49]	×1	ResNet-50	11.0	49.3	-	10.4	7.3
SML [43]	×1	ResNet-50	10.6	-	-	9.6	7.0
MGAN+OR-CNN [17]	×1	VGG16	10.5	-	47.2	-	-
	×1.3	VGG16	**9.9**	-	45.4	-	-
SMPD [ours]	×1	ResNet-50	**9.9**	**45.6**	**36.6**	**9.0**	6.5
	×1.3	ResNet-50	**9.1**	**45.9**	**36.6**	**7.9**	6.5

Table 5: Comparison with the state of the art on Citypersons. Scale indicates the scaling of the original image (1024x2048 on Citypersons) for the input. Red and blue represent the best and the second best results on the corresponding subset, respectively.

to its capability in exploiting rich semantic context to assess ambiguous cases instead of merely relying on the CNN features learned from images. Moreover, when tested on the ×1.3 image size, SMPD achieves a new state of the art with 9.1% MR^{-2}.

Caltech. Figure 7 shows comparisons with the state of the art on Caltech [20]. To compare the performance under different challenges, SMPD is also evaluated on three subsets, Heavy, Medium and All. Compared with CSP, SMPD achieves improvements in all subsets. On the reasonable setting, SMPD achieves 4.2% MR^{-2}, outperforming all previous state of the art. On the heavy occlu-



(a) Reasonable (b) Heavy occlusion

(c) Medium (d) All

Figure 7: Comparisons of the state of the art on three Caltech subsets: Reasonable, Heavy occlusion, Medium and All.

sion, medium object ([30-80]) and 'all' subsets, SMPD achieves 44.8%, 35.2% and 55.2% MR^{-2}, respectively. Since SMPD are trained with pseudo semantic labels in Caltech, the result indicates that the transfer learning approach is feasible, and SMPD has great robustness and the ability to facilitate detection in various scenarios. When models are pretrained on Citypersons, SMPD also achieves a new state of the art with 3.3% MR^{-2}. In addition, on the heavy occlusion, medium object and 'all' subsets, compared with CSP pretrained on Citypersons, SMPD achieves a large improvement of 5.5%, 3.0% and 3.2%, respectively. The substantial performance gain indicates the excellent ability of SMPD to detect occluded and small pedestrians again.

5. Discussion

It should be noted that the comparison on Citypersons is not entirely fair because we introduce additional semantic labels from Cityscapes for training. However, what we would like to share is that the proposed SPM and SIM are general auxiliary modules complementary to most detectors, and we do show a success that a leading detector can still be improved by 1.1% when scene semantics are fully exploited and used. What's more, we do show another success on Caltech without groundtruth semantic annotations. SMPD is able to use pseudo semantic labels generated by a pre-trained segmentation model to supervise training. In this case, although the pre-trained semantic predictor may have some issues in generalizing to different datasets, SMPD still improves the baseline detector by a clear margin on Caltech.

6. Conclusion

In this paper, we present the possibility that baseline pedestrian detectors could be further improved when effectively integrating rich detection cues contained in urban-scene semantics. On top of the CSP detector, the proposed SMPD achieves state-of-art results on the Citypersons and Caltech datasets. Moreover, SMPD performs especially well on small objects and occlusion cases, as well as effectively removes false positives. Due to the general structure of SMPD, it would be interesting to further examine its ability to improve other advanced pedestrian or object detectors. And because of the success in the urban scene, we are looking forward to the performance of the method in more other application scenarios. In addition, further integrating multimodal information besides semantics into the proposed model is also worth studying.

7. Acknowledgment

This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, and the National Nat-

ural Science Foundation of China under Grants 91646207, 61802407, 61773377, 62071466, and 62076242.

References

[1] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[2] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, 2016, pp. 21–37.

[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[6] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection?, in: European Conference on Computer Vision, Springer, 2016, pp. 443–457.

[7] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, 2017, p. 3.

[8] Z. Cai, Q. Fan, R. S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European Conference on Computer Vision, Springer, 2016, pp. 354–370.

[9] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Transactions on Multimedia 20 (4) (2018) 985–996.

[10] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, C. Shen, Repulsion loss: Detecting pedestrians in a crowd, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 7774–7783.

[11] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874–2883.

[12] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, S. Yan, Attentive contexts for object detection, IEEE Transactions on Multimedia 19 (5) (2017) 944–954.

[13] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE, 2017, pp. 2980–2988.

[14] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware cnn model, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1134–1142.

[15] G. Brazil, X. Yin, X. Liu, Illuminating pedestrians via simultaneous detection and segmentation, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 4960–4969.

[16] C. Lin, J. Lu, J. Zhou, Multi-grained deep feature learning for robust pedestrian detection, IEEE Trans. Circuits Syst. Video Technol. 29 (12) (2019) 3608–3621.

[17] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, L. Shao, Mask-guided attention network for occluded pedestrian detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 4966–4974.

[18] J. Mao, T. Xiao, Y. Jiang, Z. Cao, What can help pedestrian detection?, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 6034–6043.

[19] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, D. Lin, Hybrid task cascade for instance segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 4974–4983.

[20] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE transactions on pattern analysis and machine intelligence 34 (4) (2012) 743–761.

[21] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 936–944.

[22] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379–387.

[23] X. Wang, A. Shrivastava, A. Gupta, A-fast-rcnn: Hard positive generation via adversary for object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 3039–3048.

[24] T. Kong, A. Yao, Y. Chen, F. Sun, Hypernet: Towards accurate region proposal generation and joint object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 845–853.

[25] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.

[26] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 6517–6525.

[27] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, Dssd: Deconvolutional single shot detector, arXiv preprint arXiv:1701.06659.

[28] T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 318–327.

[29] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV, 2018, pp. 765–781.

[30] X. Zhou, J. Zhuo, P. Krähenbühl, Bottom-up object detection by grouping extreme and center points, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 850–859.

[31] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 9626–9635.

[32] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8) (2014) 1532–1545.

[33] W. Nam, P. Dollár, J. H. Han, Local decorrelation for improved pedestrian detection, in: Advances in Neural Information Processing Systems, 2014, pp. 424–432.

[34] P. Dollár, Z. Tu, P. Perona, S. J. Belongie, Integral channel features, in: British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings, 2009, pp. 1–11.

[35] J. Hosang, M. Omran, R. Benenson, B. Schiele, Taking a deeper look at pedestrians, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4073–4082.

[36] B. Yang, J. Yan, Z. Lei, S. Z. Li, Convolutional channel features, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 82–90.

[37] Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1904–1912.

[38] C. Li, X. Wang, W. Liu, Neural features for pedestrian detection, Neurocomputing 238 (2017) 420–432.

[39] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[40] F. B. Tesema, H. Wu, M. Chen, J. Lin, W. Zhu, K. Huang, Hybrid channel based pedestrian detection, Neurocomputing 389 (2020) 1–8.

[41] J. Cao, Y. Pang, S. Zhao, X. Li, High-level semantic networks for multi-scale object detection, IEEE Trans. Circuits Syst. Video Technol. 30 (10) (2020) 3372–3386.

[42] S. Zhang, X. Yang, Y. Liu, C. Xu, Asymmetric multi-stage cnns for small-scale pedestrian detection, Neurocomputing 409 (2020) 12–26.

[43] J. Wu, C. Zhou, Q. Zhang, M. Yang, J. Yuan, Self-mimic learning for small-scale pedestrian detection, in: C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, R. Zimmermann (Eds.), MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, ACM, 2020, pp. 2012–2020.

[44] R. Lu, H. Ma, Y. Wang, Semantic head enhanced pedestrian detection in a crowd, Neurocomputing 400 (2020) 343–351.

[45] S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Occlusion-aware R-CNN: detecting pedestrians in a crowd, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III, 2018, pp. 657–674.

[46] S. Liu, D. Huang, Y. Wang, Adaptive NMS: refining pedestrian detection in a crowd, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 6459–6468.

[47] X. Huang, Z. Ge, Z. Jie, O. Yoshie, NMS by representative region: Towards crowded pedestrian detection by proposal pairing, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, IEEE, 2020, pp. 10747–10756.

[48] W. Liu, S. Liao, W. Hu, X. Liang, X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV, 2018, pp. 643–659.

[49] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: A new perspective for pedestrian detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach,

CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 5187–5196.

[50] X. Du, M. El-Khamy, J. Lee, L. Davis, Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection, in: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE, 2017, pp. 953–961.

[51] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: European Conference on Computer Vision, Springer, 2014, pp. 297–312.

[52] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3150–3158.

[53] SSA-CNN: semantic self-attention CNN for pedestrian detection, CoRR abs/1902.09080. arXiv:1902.09080.

[54] Y. Tian, P. Luo, X. Wang, X. Tang, Pedestrian detection aided by deep learning semantic tasks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5079–5087.

[55] B. Sheng, Q. Hu, J. Li, W. Yang, B. Zhang, C. Sun, Filtered shallow-deep feature channels for pedestrian detection, Neurocomputing 249 (2017) 19–27.

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[57] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[58] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[59] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, CoRR abs/1606.02147.

[60] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1259–1267.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, Ieee, 2009, pp. 248–255.

[62] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 1195–1204.

[63] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. D. Newsam, A. Tao, B. Catanzaro, Improving semantic segmentation via video propagation and label relaxation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 8856–8865.

[64] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on topological line localization and temporal feature aggregation, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, 2018, pp. 554–569.

[65] C. Zhou, J. Yuan, Bi-box regression for pedestrian detection and occlusion estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 135–151.

[66] C. Zhou, M. Yang, J. Yuan, Discriminative feature transformation for occluded pedestrian detection, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 9556–9565.

[67] X. Chu, A. Zheng, X. Zhang, J. Sun, Detection in crowded scenes: One proposal, multiple predictions, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, IEEE, 2020, pp. 12211–12220.