IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 31, 2020, accepted September 19, 2020, date of publication September 30, 2020, date of current version October 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027590

Pedestrian as Points: An Improved Anchor-Free Method for Center-Based Pedestrian Detection

JIAWEI CAI[1], FEIFEI LEE[1], (Member, IEEE), SHUAI YANG[1], CHAOWEI LIN[1], HANQING CHEN[1], KOJI KOTANI[2], (Member, IEEE), AND QIU CHEN[3], (Member, IEEE)

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]Department of Intelligent Mechatronics, Akita Prefectural University, Akita 015-0055, Japan
[3]Major of Electrical Engineering and Electronics, Graduate School of Engineering, Kogakuin University, Tokyo 163-8677, Japan

Corresponding authors: Feifei Lee (feifeilee@ieee.org) and Qiu Chen (q.chen@ieee.org)

This work was supported in part by The Programme for Professor of Special Appointment (Eastern Scholar) at the Shanghai Institutions of Higher Learning, and in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant 15K00159.

ABSTRACT Although excessive proposals using traditional sliding-window methods or prevailing anchor-based techniques have been proposed to deal with deep learning-based pedestrian detection, it is still a promising yet challenging problem. In this paper, we propose a precise, flexible and thoroughly anchor-free, as well as proposal-free framework named Pedestrian-as-Points Network (PP-Net) for pedestrian detection. Specifically, we model a pedestrian as a single point, i.e., the center point of the instance, and predict the pedestrian scale at each detected center point. In order to achieve higher accuracy, we build a pyramid-like structure based on the backbone as a feature extractor to aggregate multi-level information. In addition, we construct a deep guidance module (DGM) at the top of the backbone, so that the higher-level information can be captured in the process of building a feature pyramid network (FPN) to avoid the dilution of high-level information on the top-down pathway. We further design a feature fusion unit (FFU) to fuse the fine-level features well with the coarse-level semantic information from the top-down pathway. With the only post-processing non-maximum suppression (NMS), we achieve better performance than many state-of-the-arts methods on the challenging pedestrian detection datasets.

INDEX TERMS Pedestrian detection, anchor-free, CNNs, feature pyramid network, deep semantic information.

I. INTRODUCTION

Deep neural networks (DNNs) based on the fully convolutional neural network have showed great improvements over systems relying on hand-crafted features [1]–[3] on benchmark tasks. With the rapid progress in DNNs research in recent years, it has dramatically facilitated the development of computer vision, such as object detection [4]–[6], image retrieval [7]–[9], scene recognition [10], [11], semantic segmentation [12]–[14], image classification and inpainting [15], [16], and so on. In particular, the state-of-the-art works in object detection continues to grow, including face recognition [17]–[19], pedestrian detection [20]–[22], vehicle detection [23], [24], etc. Pedestrians are one of main participants in the public transportation system, so pedestrian detection helps to realize an efficient and safe system. In the past few years, the widely-used anchor-based methods [25]–[29] have been dominant and have achieved tremendous progress.

Unfortunately, there are several drawbacks of current anchor-based approaches. First, anchor-based methods introduce additional hyper-parameters of design choices. Usually an extensive number of anchors, i.e., bounding boxes of the potential object are required to ensure a sufficiently high recall rate and a high Intersection over Union (IoU) rate with the ground-truth objects. Moreover, detectors encounter difficulties to manually design object candidates with large variations of size and aspect ratio of each anchor box. Second, the preset anchor boxes hinder the generality of detectors, that is, the designed choice based on a specific dataset is not always applicable to other datasets. Last but not least, another point that cannot be ignored is most of these anchor boxes are labelled as negative samples during training, leading to the imbalance between positive and negative samples.

To this end, anchor-free methods have been gradually increasing. Keypoint-based object detection [30]–[32] is a sort of methods that generate pedestrian bounding boxes by detecting and grouping their keypoints. CSP [33] detector, the state-of-the-art among them, uses the vanilla ResNet-50 network [34] to extract multi-level feature maps and then

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

concatenate them for predicting the center heatmaps and corresponding scale maps, i.e., detects the central points and size of the bounding boxes. CSP detector has achieved briliant accuracy on challenging CityPersons [35] with a simple design that eliminates the need of anchor boxes.

Unfortunately, when look closely at the operations of CSP detector, we find that the detection performance can be further improved. Specifically, after conduct feature extraction, the author of CSP simply fuses the multi-scale feature maps, which are from different stages of the backbone network, into a single one. The principle lays here is that shallower feature maps contain more accurate localization information, while the deeper ones are able to provide more semantic information as the receptive field has enlarged. However, large semantic gaps between feature maps from different depths are introduced by in-network feature hierarchy. On account of this inherent property, concatenating multi-depth feature maps directly harms their representational capacity for subsequent detection.

A large number of structures [36]–[38], which are helpful for alleviating the above problem, have been proposed. In that U-shape based structures [39], [40] can construct enriched feature maps via building top-down pathways upon basic network, they get a lot of attention. Thus, in this paper, we intuitively leverage the pyramidal shape of a ConvNet's feature hierarchy by creating a feature pyramid network (FPN) [40], where each level has strong semantic and localization information with regardless of scale. More specifically, we depend on the architecture that aggregates the features with low-resolution but strong semantic information and the ones with high-resolution yet weak semantic information through a top-down pathway and lateral connection. Taking a step further, different from the standard FPN, we investigate how to preferably solve the problem of multi-scale feature fusion when building each pyramidal level. In general, we incorporate a feature fusion unit (FFU) into our model to fuse features with different resolutions.

There is still a large room for refining existing feature pyramid network. First of all, as pointed out in [41], the semantic information captured by deep layers will be gradually diluted on the top-down pathway of the FPN architecture. Second, as mentioned in [42], the size of receptive field of a convolutional neural network (CNN) fails to be proportional to its layer depth. There are several kinds of approaches aim at addressing aforementioned problems, such as recurrently refining feature maps [43], [44], drawing attention mechanisms [45], [46] into FPN architectures, etc.

Inspired by PoolNet [41], we improve the vanilla feature pyramid network. We propose to adopt a deep guidance module (DGM) upon the bottom-up pathway, i.e., adding a residual unit on the top of the bottom-up pathway. Profiting from this operation, a higher-level feature map I with abundant semantic information can be obtained. Then, the captured information is transferred to feature maps at all pyramid levels by fusing I with them, respectively. Specifically, the deeper feature map contains extensive semantic information, thus alleviating the sparsity in top-down pathway of FPN.

In summary, the main contributions of this work can be highlighted as follows:

(1) We construct a structure in the shape of existing FPN based on ResNet-50 [34] network to obtain multi-scale features, which means that we can detect pedestrians in various scales. Then our newly-proposed feature fusion unit (FFU) together with the built feature pyramid network (FPN) can solve the problem of ignoring the large semantic gap between multi-layer features when directly fusing them.

(2) Based on the U-shaped architecture, we further build a novel deep guidance module (DGM) upon the bottom-up pathway, which aims to provide the location information of potential objects for layers at different feature levels. Therefore, we tackle the dilemma of information sparsity by expanding the role of deep features in U-shape based architectures.

(3) We develop a novel and unique framework called Pedestrian-as-Points Network (PP-Net) for real-time pedestrian detection, which can effectively utilize the semantic information of images at low resolution along with details at high-resolution.

(4) The anchor-free method achieves higher performance compared with state-of-the-art methods on CityPersons [35] and Caltech [47] datasets.

II. RELATED WORKS

Object detection has been extensively studied over the past few decades, and great progress has been made with the emergence of deep convolutional neural networks. Object detection algorithms can be classified into anchor-based and anchor-free detectors.

A. ANCHOR-BASED DETECTORS

Anchor-based detectors inherit and further expand the ideas from traditional sliding-window strategy [22] and proposal based detectors such as Fast R-CNN [47]. Pedestrian detection has been significantly improved due to the use of dense predefined anchors with preset scales and aspect ratios. Modern CNN-based detectors are categorized into two-stage and one-stage detectors. Within the two-stage framework, classical Faster R-CNN [26] utilizes an anchor mechanism in the branch dedicated to generating proposals, i.e., Region Proposal Network (RPN). Afterwards, dozens of methods [27], [48]–[50] have been developed.

Following Faster R-CNN, Mask R-CNN [48] adds a mask branch parallel to the branch of classification and regression for performing mask predicting. For the sake of preventing Faster R-CNN from heavy region-wise CNN computational cost, R-FCN [49] proposed efficient region-wise fully convolutions without accuracy loss. Cascade R-CNN [27] extends the architecture of Faster R-CNN to multiple stages. Illumination-aware faster R-CNN [50] addresses the problem of fusing color and thermal modalities for detecting multispectral images. In the one-stage stream, a considerable number of approaches [51]–[55] which use anchor mechanism are proposed after SSD [25]. They aim at improving performance, including multi-stage refinement [51], [52], adaptive anchors [53] and loss function improvement [54], [55].

B. ANCHOR-FREE DETECTORS

Most recently, a lot of papers about anchor-free [56]–[60] have published, which has a great momentum of transforming the period of anchor-based detector.

CornerNet [56] predicts two groups of corners of bounding box, i.e., top-left and bottom-right points and then divides the corners belonging to the same object into a group based on the distance between the corner embedding by Grouping Corners, which is inspired by the Associative Embedding method [57]. Corner pooling is used for better localizing the corners. CornerNet-Lite [58] is a combination of two efficient variants of the CornerNet and thus improves efficiency without sacrificing accuracy. Compared to CornerNet, the ExtremeNet [59] detects four extreme points and central points of bounding box instead of corners. More specifically, top, left, bottom and right points are predicted and then grouped to form the final detected bounding box. FCOS [60] predicts the bounding boxes by making full use of the advantages of all points in a ground truth bounding box. And the low-quality detected bounding boxes are suppressed by the proposed "center-ness" branch. The detector considers location of object as training sample rather than anchor box, which is same as semantic segmentation.

Following anchor-free pipeline, our work aims to predict the precise center points and the corresponding pedestrian scales. We try to explore whether the results of such a simple method of localizing pedestrians by simply detecting the center points can be more competitive than other complex methods.

C. FEATURE PYRAMID NETWORKS

Feature pyramid constructing module are applied many computer vision applications required multiscale processing as the basis of solutions. Furthermore, the feature pyramid representation module can be easily modified and insert into most deep neural networks based detectors. SPPNet [61] eliminates ConvNet's requirements for fixed input by introducing spatial pyramid pooling layer. Recently, PFPNet [62] extends the idea to build multiple parallel SPPNets for generating feature pools with different sizes, then the elements in the feature pool are rescaled to a uniform size and their context information aggregated to generate each level of the final feature pyramid. M2det [63] employs multiple U-shaped modules after a backbone model and thus build stronger feature pyramid representations. NAS-FPN [64] introduces the Neural Architecture Search (NAS) mechanism and discovers the fresh feature pyramid architecture in a novel scalable search space covering all cross-scale connections.

D. PEDESTRIAN DETECTION

As a critical part of general object detection, pedestrian detection receives considerable interests. Nowadays, the field of pedestrian detection is almost dominated by deep learning [28], [65]–[67].

A jointly learning framework is proposed by [65]. In addition, [63] adds extra features to improve performance. A cascaded prediction is performed by [28] to stimulate the potential of one-stage detectors. [66], [67] focus on studying and overcoming the impact of occlusion. [66] is the first one operates full and visually body of a pedestrian regression simultaneously. Reference [67] employs attention mechanism into framework for enhancing the features of pedestrian.

In our work, we aim at putting up with multi-scale problems of pedestrian detection by refining feature pyramid network [68]–[71]. Reference [68] enhances the semantic information of low-level features by applying multiple convolution operations and increases resolution of high-level features by getting rid of pooling layer. Reference [69] applies feature pyramid network, equipped with refined attention modules to strengthen the representation ability of features. Reference [70] enhances feature pyramid network by introducing a cross-scale feature aggregation module. In [71], the structure of convolution neural network is summarized.

III. PROPOSED METHOD

As is known to all, high-level features with rich semantic information help to discover specific locations of objects. Meanwhile, low- and mid-level features with plenty of location information are also essential for refining the coarse features extracted from deep layers. Based on the above knowledge, we propose in this section a pyramid-like network named Pedestrian-as-Points Network (PP-Net) as illustrated in Fig. 1, which has two complementary modules that can detect the exact positions of pedestrians and simultaneously predict their sizes.

A. OVERALL PIPELINE

We build our architecture in a one-stage manner. It takes advantages of the widely-adopted feature pyramid network (FPN) [40], which is a kind of classic U-shaped architectures designed in a bottom-up and top-down manner for finely fusing multi-level features as shown in Fig. 4a.

For the bottom-up pathway, we adopt ResNet-50 [34] as the forward baseline network unless otherwise stated, which consists of five stages made up with Conv layers (convolutional layer followed by batch normalization and ReLU). It is worth noting that feature tensors with the same scale belong to a network stage. It is natural for us to choose the last feature map of each stage as our reference set of feature maps, which we will enrich to generate our pyramid-like structure, because the feature map with the strongest representation ought to exist in the deepest layer of each stage. The output feature maps of different stages in the forward streamline are down-sampled by 2, 4, 8, 16, 32 w.r.t. the input image.

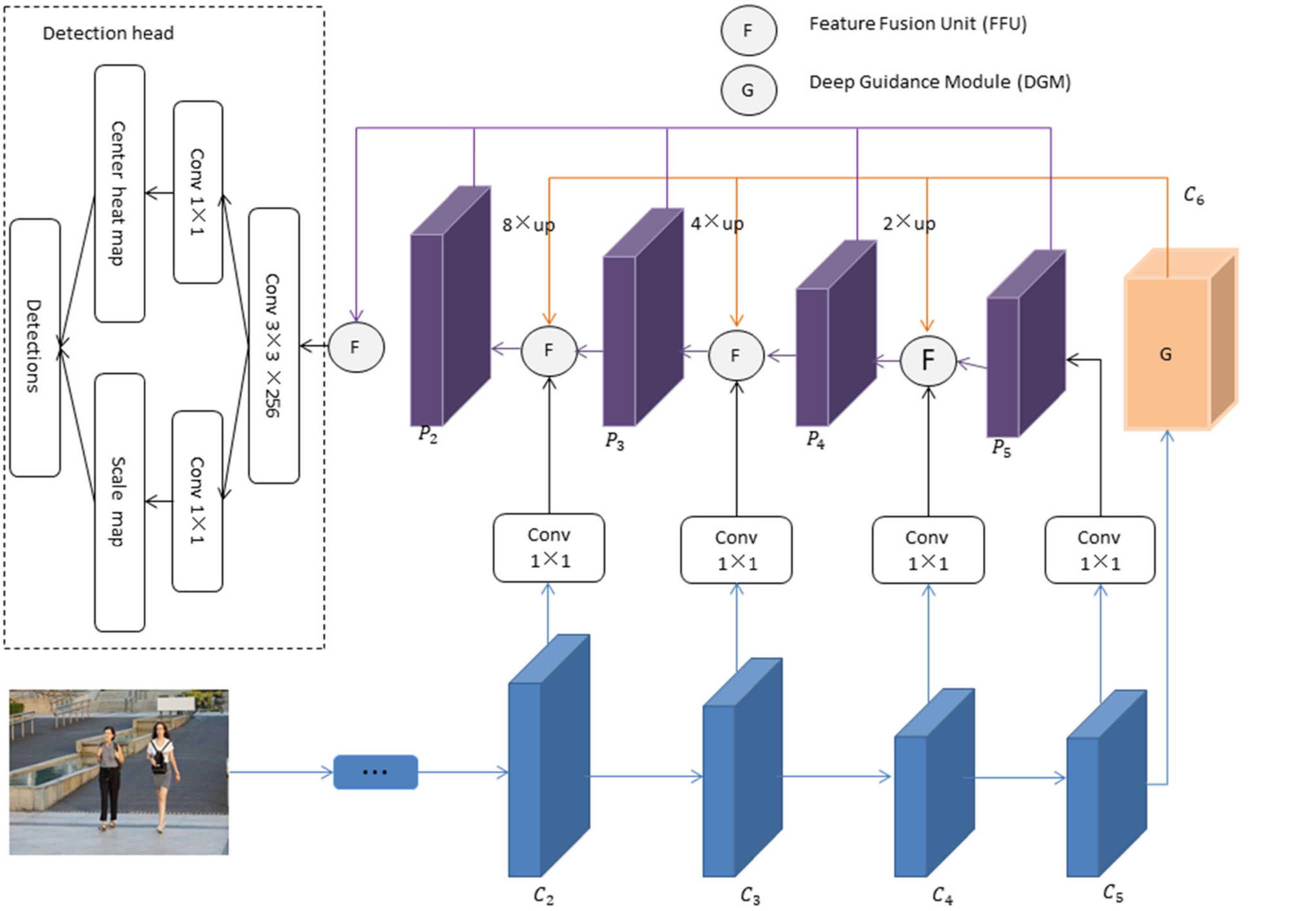


FIGURE 1. The overview of our proposed framework named Pedestrian-as-Points Network (PP-Net). PP-Net models a pedestrian as a single point, i.e., the center point of the instance, and predicts the pedestrian scale at each detected center point. A pyramid-like structure is built to aggregate multi-level information. The deep guidance module at the top of the backbone can provide the higher-level information, thereby avoiding the dilution of high-level information on the top-down pathway. The proposed feature fusion unit fuses fine-level features well with coarse-level semantic information from the top-down pathway.

In practice, the output of stage 5 is kept as 1/16 of the input image size by utilizing the dilated convolutions. We denote the last feature map of each stage as C_i, where i corresponds to the stage within the backbone hierarchy. Concretely, the last feature maps of stage 2, 3, 4 and 5 are denoted as C_2, C_3, C_4 and C_5, in which the shallower feature maps contain more accurate localization information, while the deeper ones can provide more semantic information with larger receptive fields. We do not include stage 1 into the building of pyramid due to its large memory footprint.

As shown in the Fig. 1, we add a deep guidance module (DGM) to address the feature dilution on the top of the bottom-up pathway. More specifically, we explicitly transform the guidance information from DGM to the layers at different feature levels by merging the high-level information extracted by DGM with feature maps at each feature level. After then, we actually go one step further and introduce a feature fusion unit (FFU) to ensure that feature maps at different resolutions can be concatenated seamlessly.

The features with higher resolution are generated in top-down pathway via up-sampling spatially coarser yet semantically stronger feature maps from higher pyramid levels. For the efficient design of FPN, we aim to make the pyramid pathways lightweight by reducing their channel capacity. To be specific, the channel capacity which is significantly lower than the number of channels of the final stage in the backbone pathway is used, yielding the computationally-effective multiple pathways because the computation cost of a weight layer scales quadratically with its channel dimensions.

In detail, we first attach a 1 × 1 convolutional layer on C_5 to produce the coarsest resolution map $C_5{}'$. Here, the 1 × 1 convolutional layer is used for reduce channel dimensions to fixed number, denoted as d (d = 256 in the paper). Then, the feature maps $C_5{}'$ and C_6 (output of GMM) are fed into FFU, creating a feature map P_5. Then we reduce the number of channels of C_4 to d via a 1 × 1 convolutional layer and feed the output along with P_5 and C_6 into FFU for generating P_4. This process is iterated until the finest resolution map P_2 is obtained. It is noteworthy that if the resolution among the inputs of FFU is different, we are going to rescale the coarser ones by apply up-sampling rate 2 on them through bilinear interpolation. Last but not least, the number of channels is reduced to d by applying 1 × 1 convolution operation before sent into the feature fusion module.

Finally, we append a detection head, which is crucial in the whole detection system, to the generated feature map P_2 to parse it into the final detection results. The structure of the head is shown in the Fig.1. First, the number of channels is reduced to 256 by a 3×3 convolutional layer. Then, the center heatmap and scale map are produced separately via two parallel 1×1 convolutional layers. The predicted heatmaps are with the same size as the concatenated feature maps. Note that more complicated detection head like [52], [55] can be explored to further improve the detection performance, but it beyond the scope of this work.

The following two reasons can explain why anchor-free detection is superior to anchor-based one, i.e., why detecting centers is more effective than bounding box proposals. First, from CornerNet [56] we can know that directly predicting the center points is a more efficient way for densely discretizing the space of boxes, because $\mathrm{O}\left(\mathrm{w}^2\mathrm{h}^2\right)$ possible anchor boxes can be represented by only O(wh) centers. Second, the anchor-free way has a smoother prediction, which can empirically improve the generalization performance of the network. Third, the anchor-free method avoids a large amount of IOU calculation between GT boxes and anchor boxes, so that the training process takes up less memory.

Subsequently, we will describe the architectures of the two modules mentioned above, namely Deep Guidance Module (DGM) (Sec.3.2) and Feature Fusion Unit (FFU) (Sec.3.3), and describe their functions in detail.

B. DEEP GUIDANCE MODULE

There are two main noticeable issues caused by constructing top-down pathway of U-shaped structure based on the bottom-up backbone. One of them is the dilution of the deep semantic information in the top-down transportation way. The other is the misalignment between receptive field in practice and theory. In particular, it is not sufficient for the small virtual receptive field of the CNNs to cover the entire input images. To this end, we propose a deep guidance module (DGM) for providing deeper and richer information, which is in a plug-and-play manner.

As shown in Fig.2 (c), the structure of the DGM is adapted from the residual stage of original ResNet. Inspired by DetNet [72], our proposed deep guidance module consists of a dilated bottleneck with 1×1 convolution projection and two subsequent dilated bottleneck identical connection. To be more specific, as shown in the Fig.2a and Fig.2b, we apply bottleneck with dilation as a basic unit of DGM for efficiently exploring deep semantic information while enlarging the receptive filed without changing fixed spatial size of feature map after stage 5.

C. FEATURE FUSION UNIT

In order to fuse feature maps with different resolutions for constructing feature pyramid structure, we propose a simple while effective feature fusion module. As shown in Fig.3, the inputs of feature fusion module are three feature maps with different scales. More precisely, they represent the feature maps fused to build a pyramidal level in the top-down pathway of our new structure, i.e., feature maps F_1,F_2 and F_3 with sizes $C_1{\times}H_1 \times W_1$, $C_2{\times}H_2 \times W_2$ and $C_3{\times}H_3 \times W_3$. Note that F_3 is with doubled spatial size of F_1 and F_2. In other words, the resolution of F_1 is equal to the one of feature map F_2.

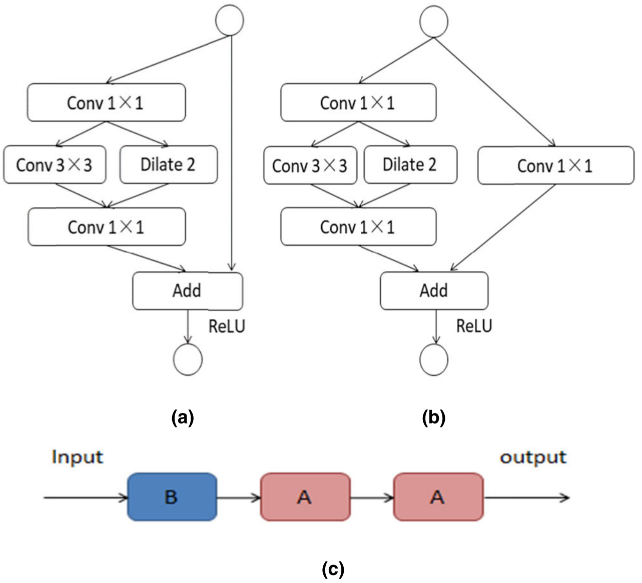


FIGURE 2. (a) The overview of Dilated bottleneck, (b) The overall structure of Dilated bottleneck with 1 × 1 convolution projection, (c) Architecture of our proposed deep guidance module (DGM), where A denotes Dilated bottleneck and B is Dilated bottleneck with 1 × 1 convolution projection.

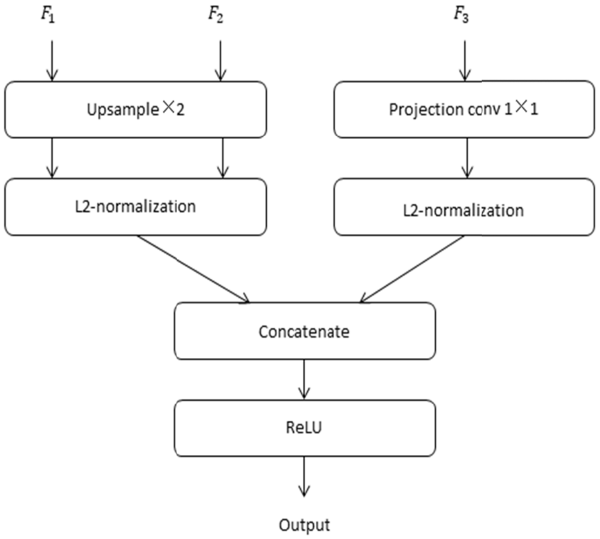


FIGURE 3. The whole architecture of developed feature fusion unit (FFU).

For $F_1(F_2)$, we first double the resolution of $F_1(F_2)$ via a deconvolution layer, leading to the same size as feature map F_3. Then a L2-normalization layer is used to rescale the norm of the resized feature map for following fusion operation.

As for F_3, since there is no necessity of changing spatial size, we merely carry out the L2-normalization for adjusting the norm to the same as the one of processed F_1.

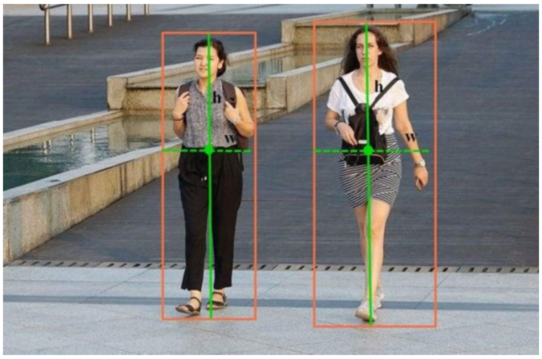


FIGURE 4. Illustration of generation process of bounding box. The green dots denote the central point of each pedestrian. The green line ***h*** represents the height of the pedestrian. The green dotted line ***w*** is the corresponding width of the pedestrian. The orange box is the final bounding box.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we first describe the details of training the framework. Then, we introduce the implementation details, the used datasets and the evaluation metrics of the experiment. Next, we exhibit the experiment results and the comparison among previous state-of-the-art methods. Finally, we demonstrate the effectiveness of each module we proposed through a series of ablation studies.

A. TRAINING

We can generate ground truth map of center and scale with the bounding box annotations. For the center ground truth, if a location is the center point of a pedestrian, it is defined as positive, and vice versa.

As for scale, it can be defined as the height and width of pedestrians. Following the CSP detector [33], we only predict the height of each pedestrian, and then the bounding box can be obtained by the preset aspect ratio because we define that the high-quality ground-truth bounding boxes are automatically generated by a uniform aspect ratio of 0.41. Additionally, the values of $\log(h_k)$ corresponding to the k-th object are allocated to the k-th positive locations and the negatives within a radius 2 of the positives (for alleviating ambiguity), while all other locations are assigned as zero. Specifically, the framework directly predicts a 1D vector, i.e., the height information of the object plus a class category at each positive location on a level of feature maps. As shown in Fig. 4, the four sides of a bounding box (shown as orange box in the figure) can be obtained through 1D vector (shown as vertical green solid line).

We adopt the classification loss in [31] which can be formulated as:

$$L_{cls} = -\frac{1}{K}\sum_{i=1}^{\frac{W}{r}}\sum_{j=1}^{\frac{H}{r}}\alpha_{ij}(1-\widehat{p_{ij}})^{\gamma}\log(\widehat{p_{ij}}) \tag{1}$$

where

$$\widehat{p_{ij}} = \begin{cases} p_{ij}, & if\ y_{ij}=1 \\ 1-p_{ij}, & otherwise, \end{cases} \tag{2}$$

$$\alpha_{ij} = \begin{cases} 1, & if\ y_{ij}=1 \\ (1-M_{ij})^{\beta}, & otherwise, \end{cases} \tag{3}$$

In the equation, $\mathrm{p_{ij}} = 1$ if the center point of object pedestrian is located in the coordinate (i, j) while otherwise 0. And $\mathrm{y_{ij}} \in \{0, 1\}$ denote the ground truth label. In addition, M is the mask map and is calculated by using a 2D Gaussian mask $G(\cdot)$, which is proposed for relieving the ambiguity of negative samples surrounding the positive ones, it is formulated as in CSP detector [33].

On the whole, the final loss function is:

$$\mathrm{L} = \lambda_{\mathrm{cls}}\mathrm{L}_{\mathrm{cls}} + \lambda_{\mathrm{reg}}\mathrm{L}_{\mathrm{reg}} + \lambda_{\mathrm{o}}\mathrm{L}_{\mathrm{offset}} \tag{4}$$

Here, $\mathrm{L}_{\mathrm{reg}}$ and $\mathrm{L}_{\mathrm{offset}}$ are both adapted from the smooth L1 loss function.

B. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

Our proposed framework is implemented in Keras [73]. Totally, our network is trained for 150 epochs in total and the optimizer is Adam [74] with an initial learning rate of 2e-4. By default, the backbone is pre-trained ResNet-50 and the rest modules are randomly initialized. During the test phase, we extract the results from the models trained with 50 to 150 epochs respectively unless otherwise stated.

2) DATASETS

For verifying the availability of PP-Net, we use two challenging datasets CityPersons [35] and Caltech [47], which can provide central point annotations and aspect ratios of bounding boxes. CityPersons contains 2975 images for training and 500 images for testing, to demonstrate the performance of proposed framework. The images of CityPersons are in extremely large sizes and the types of occlusions are many and varied. Caltech consists of 42782 training images and 4024 testing images, which are the frames extracted from a 2.5-hour auto-driving video.

Compared to other datasets, the annotations of these two selected datasets highly fit for our method as they contain normalized aspect ratio and central body line annotation.

Before training, some methods of data augment are used, such as random brightness, random crop and color jittering.

3) METRICS

Follow the CSP detector [33], we choose the log-average miss rate against false positives per image (MR-FPPI) (ranging in $[10^{-2}, 10^{0}]$), which we denote as MR, for evaluating the detection results. The calculation of the miss rate can be seen in [22]. Also, we use average precision (AP) for supplement. Note that the higher the value of average precision (AP), the higher the accuracy of the pedestrian detected by the detector. While the value of miss rate (MR) is about low, which means that the number of pedestrians missed by the detectors is less.

C. COMPARISON WITH STATS-OF-THE-ART METHODS

1) CityPersons

In this section, we compare our proposed framework with several previous state-of-the-art methods in CityPersons dataset [35], including FRCNN [35], FRCNN+Seg [35],

TABLE 1. Comparison with state-of-the-art methods on CityPersons.

Method	Miss Rate		AP
	Reasonable	Heavy	
FRCNN+Seg[35]	14.80%	–	93.80%
TLL+MRF[77]	14.40%	52.0%	–
RepLoss[76]	13.20%	56.9%	94.70%
AdaptiveNMS[78]	12.90%	56.4%	95.30%
OR-CNN[75]	12.80%	55.7%	–
PBM+Mask[79]	12.30%	53.3%	–
DCS+NMS[6]	11.70%	–	95.20%
CSP[33]	14.02%	56.9%	90.47%
PP-Net	12.12%	52.9%	93.78%

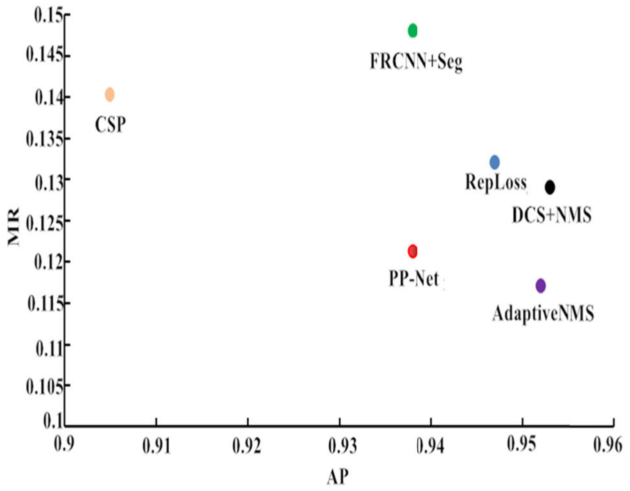


FIGURE 5. Comparison with the state of the arts on CityPersons. MR denotes miss rate on reasonable set.

OR-CNN [75], RepLoss [76], TLL+MRF [77] and CSP detector [33]. For fair comparisons, the final detection results of aforementioned methods are directly provided by authors except our closest competitor CSP detector, i.e., CSP detector is re-implemented by the original code released by the authors with Keras [73]. In the table I, it is can be found that our proposed method (denoted as PP-Net in the table) outperforms most methods above, especially main comparison object CSP detector. In other words, we reach a competitive performance of pedestrian detection in the challenging dataset in spite of the various occlusions and scales.

Moreover, as illustrated in Fig.5, form the horizontal axis, PP-Net performs barely satisfactory. Fortunately, it is close to the number one DCS+NMS [6]. That is to say, the AP of PP-Net is just passable. From the vertical axis, PP-Net performs well and is superior to most methods.

In brief, our proposed PP-Net combines accuracy with strong object capture capability.

Several qualitative results are shown in Fig. 6. It indicates that our proposed PP-Net can detect great majority of pedestrians even some of them are crowded, highly overlapped, small and large.

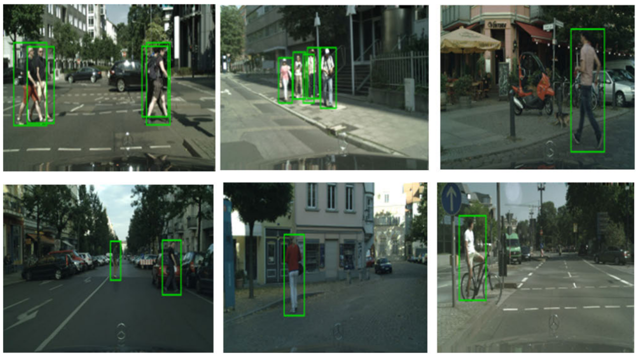


FIGURE 6. Several detection results on CityPersons.

TABLE 2. Comparison with state-of-the-art methods on Caltech.

Thresholds	0.40	0.45	0.50	0.55	0.60
MR(CSP[33])	7.38%	7.38%	10.37%	13.14%	17.58%
MR(PP-Net)	7.45%	8.13%	9.54%	12.56%	17.24%

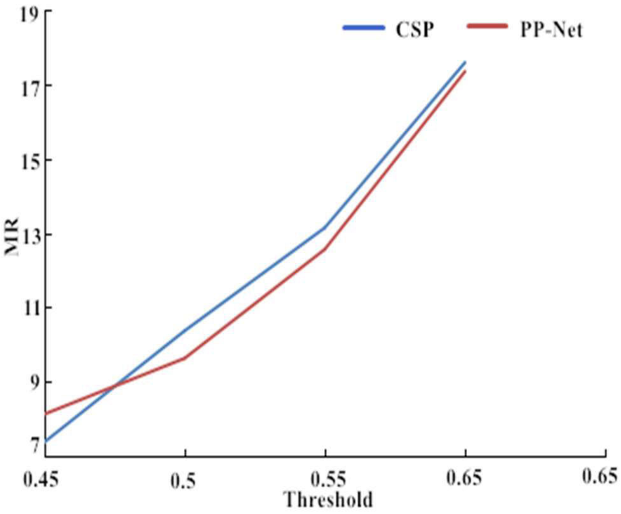


FIGURE 7. Curves of miss rate across various NMS thresholds.

2) CALTECH

Table 2 and corresponding Fig.7 show the comparisons with state of the arts on Reasonable setting across multiple NMS thresholds. We also re-implement the CSP detector [33] for the sake of fairness. Our PP-Net achieves passable result, which is comparable with the main competitor CSP detector. Because there are not sufficient training samples for the model to be fully trained, the improvement is slightly inferior to that on the CityPersons dataset. In PP-Net, each prediction point is not associated with a particular reference shape, and it directly predicts the bounding boxes with the predicted height information. Since PP-Net allows specific aspect ratios, it can capture the entire body of a pedestrian in a similar shape.

From Fig.7, we can draw a conclusion that our PP-Net is less sensitive to the NMS thresholds because its curve is smoother than that of baseline.

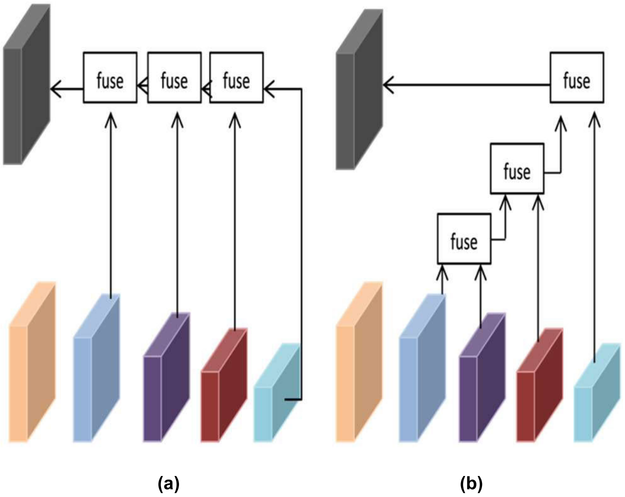


FIGURE 8. **(a) Standard U-shaped FPN structure [25] (b) The overview of Nearest-fused architecture proposed for reducing semantic gaps among different levels. Fuse-nearest denotes the result of our proposed structure in Fig.4b.**

D. ABLATION STUDY

In this subsection, we demonstrate the effectiveness of different components which we introduce in our proposed framework with different settings. To reach the goal, we construct several variants and evaluate them on convincing CityPersons [35] and Caltech [47] datasets.

1) U-SHAPED STRUCTURE

We put to use U-shaped structure upon the basic ResNet-50 for narrowing the semantic gaps between different-depth features. Meanwhile, we also design another alternative structure for the same purpose. As is shown in the Fig.8b, we gradually fuse the feature maps from the nearest two stages instead of directly fusing all feature maps, termed as Nearest-fused architecture. The results on different datasets are compared as displayed in Table 3 (a) and (b) respectively. Besides, we directly construct these two feature fusion structures on the backbone network of initial CSP detector, so as to eliminate the influence of the deep guidance module. And the results comparisons with various datasets are shown in Table 3 (c) and (b).

From the results above, it is suggested that the proposed alternative architecture is inferior to U-shaped FPN structure. We can see from Table 3 (a) and III (c) that on CityPersons dataset, the U-shaped structure improves baseline method by reducing the miss rates (MR) by 0.69% and 1.15% with and without deep guidance module respectively while Nearest-fused architecture only correspondingly reduces by 0.64% and 0.89%, which demonstrates the effectiveness of FPN.

In addition, from the Table 3 (b) and III (d), it can be observed that on Caltech dataset, the U-shaped structure improves baseline method by reducing miss rates (MR) by 0.66% and 0.16% with and without deep guidance module respectively while Nearest-fused architecture hurts the performance.

TABLE 3. **(A) Ablation study of U-shaped structure with DGM on CityPersons. (B) Ablation study of U-shaped structure without DGM on Caltech. (C) Ablation study of U-shaped structure with DGM on CityPersons. (D) Ablation study of U-shaped structure without DGM on Caltech.**

(a)

Structure	Miss Rate
Baseline	12.81%
Nearest-Fused	12.17%
U-shape	12.12%

(b)

Structure	Miss Rate
Baseline	10.20%
Nearest-Fused	10.22%
U-shaped	9.54%

(c)

Structure	Miss Rate
Baseline	14.02%
Nearest-Fused	13.13%
U-shaped	12.87%

(d)

Structure	Miss Rate
Baseline	10.37%
Nearest-Fused	12.68%
U-shaped	10.21%

2) DEEP GUIDANCE MODEL (DGM)

For proving the performance improvement brought by the proposed deep guidance module (DGM), we add DGM upon the backbone of the feature extraction part of CSP detector [33]. We then concatenate the feature maps from stage 3, 4, 5 and DGM for following detection. For verification, we take DGM away from our proposed framework and test the performance (Note that we only detect the feature map from the bottom level of FPN-like network for simplicity). The result on two datasets are shown in the Table 4 (a) and (b). We can observe that DGM plays an important role in our detector.

For further exploration, we employ atrous spatial pyramid pooling (ASPP) from DeepLab V3 [13] to substitute the original deep guidance module (DGM). The ASPP consists of several parallel branches of atrous convolution with different dilated rates to capture multi-scale context. Following the configurations in [13], ASPP consists of one 1×1 convolution and three 3×3 convolutions with rates $= (6, 12, 18)$ when output stride $= 16$ (all with 256 filters and batch normalization), and the image-level feature obtained by operating a

TABLE 4. (A) Ablation study of deep guidance model on CityPersons. (B) Ablation study of deep guidance model on Caltech.

(a)

Deep Guidance Model	U-shape Structure	Miss Rate [a]
		14.02%
√		12.81%
	√	12.87%
√	√	12.80%

(b)

Deep Guidance Model	U-shape Structure	Miss Rate [a]
		10.37%
√		10.20%
	√	10.21%
√	√	9.94%

TABLE 5. (A) Comparison between different DGM with FPN on CityPersons. (B) Comparison between different DGM with FPN on Caltech. (C) Comparison between different DGM without FPN on CityPersons. (D) Comparison between different DGM without FPN on Caltech.

(a)

Structure	Miss Rate
Baseline	13.47%
DGM	12.12%
ASPP	13.69%

(b)

Structure	Miss Rate
Baseline	10.21%
DGM	9.54%
ASPP	9.65%

(c)

Structure	Miss Rate
Baseline	14.02%
DGM	12.81%
ASPP	14.12%

(d)

Structure	Miss Rate
Baseline	10.37%
DGM	10.20%
ASPP	11.28%

1 × 1 convolution with 256 filters on the last feature map of the model. The dilated convolutions with same kernel size and different dilated rate possess different receptive field. In terms of previous work and theory, ASPP can provide multi-scale representation with deep information if it replaces our deep guidance module (DGM).

The result comparisons on two datasets are shown in Table 5 (a) and (b). On CityPersons, it can be found from the table that our DGM brings about reduction of 1.35% in MR while ASPP promotes by 0.22%, which means our DGM is able to provide more semantic information beneficial for the final prediction while operating multi-branch dilated convolutions on final feature maps may generate redundant feature information which greatly disturbs the detection results. On Caltech dataset, our DGM reduces MR by 0.67%, while ASPP reduces by 0.58%, showing that both ASPP and our DGM help to improve results, but our DGM brings more growth.

To further test and verify our point and remove interference brought by feature fusion architecture (i.e., U-shaped FPN structure), we conduct experiment on vanilla CSP detector without follow-up FPN. As in Table 5 (c) and (d), it is suggested that on CityPersons, our DGM improves MR of the baseline method by 1.21% while ASPP degrades the performance instead, which implies that not all modules can provide semantic information that is helpful for detection performance.

In addition, on Caltech, we can also draw similar conclusion that our DGM is helpful to improve the vanilla CSP detector [33] with 0.17% reduction of MR, while the ASPP has a negative effect.

3) FEATURE FUSION UNIT (FFU)

We consider that our feature fusion module is superior to previous operation which fuses multi-scale feature maps directly. To this end, we conduct the removal of FFU module.

The numerical results in Table 6 (a) and (b) indicate that the absence of FFU module is harmful for the performance of our approach by increasing MR by 0.45% on CityPersons and 0.09% on Caltech because the various norms of multi-scale feature maps play a negative role in the process of feature fusion. Compared with the existing feature fusion module, our FFU is simple and pragmatic.

4) AGGREGATE ALL LEVELS OR NOT

While building the U-shaped structure, we intuitively face with two related choices. More concretely, which level of the structure is the feature with finest resolution, i.e., the bottom level should be P_2 or P_3 ?

The other one is whether we should detect the feature maps via fusing all levels of FPN-like network or the one from the bottom level? In order to make the most beneficial decision to improve performance, we conduct the comparison experiments and the results on two datasets can be seen in the Table 7 (a) and (b) separately.

It is demonstrated that we should build P_2 as the bottom level in the top-down pathway, and concatenate all levels of U-shaped network for subsequent detection.

TABLE 6. (A) Ablation study of feature fusion unit on CityPersons. (B) Ablation study of feature fusion unit on Calech.

(a)

Structure	Miss Rate
Baseline	14.02%
PP-Net(w/o FFU)	12.57%
PP-Net	12.12%

(b)

Structure	Miss Rate
Baseline	10.37%
PP-Net(w/o FFU)	9.63%
PP-Net	9.54%

TABLE 7. (A) Comparison of aggregate all levels and only bottom level of FPN on CityPersons. (B) Comparison of aggregate all levels and only bottom level of FPN on Caltech.

(a)

p_2	p_3	Aggregation	Miss Rate [a]
√			12.80%
√		√	12.12%
	√		12.85%
	√	√	12.87%

(b)

p_2	p_3	Aggregation	Miss Rate [a]
√			9.94%
√		√	9.54%
	√		15.08%
	√	√	14.27%

Note that the definition of symbols in table are same as Table VII (a).

V. CONCLUSION

In this paper, we have proposed an anchor-free pedestrian detector which finds a better trade-off between accuracy and efficiency. We have established a U-shaped architecture to eliminate the semantic gaps between multi-level feature maps. In addition, we propose a deep guidance module to extract deep semantic information for addressing the information dilution in the top-down pathway. We further propose a feature fusion unit (FFU) for multi-feature concatenation. By plugging these modules into the FPN-like network, we can achieve significant performance. The detection results on the challenging CityPersons and Caltech datasets demonstrate that our framework is competitive with the state-of-the-art methods. In our future work, we will pursue better performance by exploring superior detection heads.

ACKNOWLEDGMENT

(Jiawei Cai and Feifei Lee contributed equally to this work.)

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[4] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "ORSIm detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jul. 2019.

[5] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, Feb. 2020.

[6] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2020, pp. 12214–12223.

[7] H. You, S. Tian, L. Yu, and Y. Lv, "Pixel-level remote sensing image recognition based on bidirectional word vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.

[8] X. Wang, F. Lee, and Q. Chen, "Similarity-preserving hashing based on deep neural networks for large-scale image retrieval," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 260–271, May 2019.

[9] H. Shojanazeri, D. Zhang, S. Wei Teng, S. Aryal, and G. Lu, "A novel perceptual dissimilarity measure for image retrieval," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2018, pp. 1–6.

[10] X. Zeng, Y. Zhang, X. Wang, K. Chen, D. Li, and W. Yang, "Fine-grained image retrieval via piecewise cross entropy loss," *Imag. Vis. Comput.*, vol. 93, Jan. 2020, Art. no. 103820.

[11] L. Xie, F. Lee, L. Liu, Z. Yin, and Q. Chen, "Hierarchical coding of convolutional features for scene recognition," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1182–1192, May 2020.

[12] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107205.

[13] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[14] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[15] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[16] W. Cai and Z. Wei, "PiiGAN: Generative adversarial networks for pluralistic image inpainting," *IEEE Access*, vol. 8, pp. 48451–48463, 2020.

[17] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.

[18] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5325–5334.

[19] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1522–1530.

[20] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[21] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[22] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[23] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.

[24] Y. Zhou, L. Liu, L. Shao, and M. Mellor, "DAVE: A unified framework for fast vehicle detection and annotation," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Aug. 2016, pp. 278–293.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Dec. 2016, pp. 21–37.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[27] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[28] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Sep. 2018, pp. 618–634.

[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[30] P. Gao, R. Yuan, F. Wang, L. Xiao, H. Fujita, and Y. Zhang, "Siamese attentional keypoint network for high performance visual tracking," 2019, *arXiv:1904.10128*. [Online]. Available: http://arxiv.org/abs/1904.10128

[31] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[32] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[33] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Feb. 2017, pp. 3213–3221.

[36] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[37] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3127–3135.

[38] Q. Hou, J.-J. Liu, M.-M. Cheng, A. Borji, and P. H. S. Torr, "Three birds one stone: A general architecture for salient object segmentation, edge detection and skeleton extraction," 2018, *arXiv:1803.09860*. [Online]. Available: http://arxiv.org/abs/1803.09860

[39] O. Ronneberger, P. Fischer, and T. Brox, "UNet: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, May 2015, pp. 234–241.

[40] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3127–3135.

[41] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3912–3921.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[43] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[44] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency Detection with Recurrent Fully Convolutional Networks," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Sep. 2016, pp. 825–841.

[45] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 714–722.

[46] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3089–3098.

[47] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, Dec. 2015, pp. 1440–1448.

[48] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, Mar. 2017, pp. 2980–2988.

[49] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[50] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, Jan. 2020.

[51] X. Wu, D. Zhang, J. Zhu, and S. C. H. Hoi, "Single-shot bidirectional pyramid networks for high-quality object detection," 2018, *arXiv:1803.08208*. [Online]. Available: http://arxiv.org/abs/1803.08208

[52] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, Jun. 2018, pp. 4203–4212.

[53] T. Yang, X. Zhang, Z. Li, W. Zhang, and J. Sun, "Meta Anchor: Learning to detect objects with customized anchors," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 318–328.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Aug. 2017, pp. 2999–3007.

[55] K. Chen, J. Li, W. Lin, J. See, J. Wang, L. Duan, Z. Chen, C. He, and J. Zou, "Towards accurate one-stage object detection with AP-loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5114–5122.

[56] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Aug. 2018, pp. 734–750.

[57] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 2274–2284.

[58] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-lite: Efficient keypoint based object detection," 2019, *arXiv:1904.08900*. [Online]. Available: http://arxiv.org/abs/1904.08900

[59] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.

[60] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, Nov. 2019, pp. 9626–9635.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[62] S. Kim, H. Kook, J. Sun, M. Kang, and S. Ko, "Parallel feature pyramid network for object detection," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Sep. 2018, pp. 234–250.

[63] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2DET: A single-shot object detector based on multi-level feature pyramid network," in *Proc. AAAI*, Feb. 2019, pp. 9259–9266.

[64] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7029–7038.

[65] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, May 2017, pp. 3127–3136.

[66] C. Zhou and J. Yuan, "Bi-box regression for pedeatrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Sep. 2018, pp. 135–151.

[67] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[68] J. Cao, Y. Pang, and X. Li, "Exploring multi-branch and high-level semantic networks for improving pedestrian detection," 2018, *arXiv:1804.00872*. [Online]. Available: http://arxiv.org/abs/1804.00872

[69] Z. Chen, L. Zhang, A. M. Khattak, W. Gao, and M. Wang, "Deep feature fusion by competitive attention for pedestrian detection," *IEEE Access*, vol. 7, pp. 21981–21989, 2019.

[70] X. Zhang, S. Cao, and C. Chen, "Scale-aware hierarchical detection network for pedestrian detection," *IEEE Access*, vol. 8, pp. 94429–94439, 2020.

[71] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, "Taking a look at small-scale pedestrians and occluded pedestrians," *IEEE Trans. Image Process.*, vol. 29, pp. 3143–3152, 2020.

[72] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: A backbone network for object detection," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Apr. 2018, pp. 334–350.

[73] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[74] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015.

[75] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput Vis. (ECCV)*, Jul. 2018, pp. 657–674.

[76] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "RepulsionLoss: Detecting pedestrians in A crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, Nov. 2018, pp. 7774–7783.

[77] T. Song, L. Y. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 643–659.

[78] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6452–6461.

[79] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020.

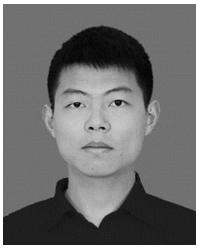
[80] T. Yin, X. Zhou, and K. Philipp, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," 2020, *arXiv:2006.11275*. [Online]. Available: http://arxiv.org/abs/2006.11275

JIAWEI CAI received the B.S. degree in detection guidance and control technology from Shenyang Aerospace University, Shenyang, China, in 2018. He is currently pursuing the M.S. degree in control science and control engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and deep learning.

FEIFEI LEE (Member, IEEE) received the Ph.D. degree in electronic engineering from Tohoku University, Japan, in 2007. She is currently a Professor with the University of Shanghai for Science and Technology. Her research interests include pattern recognition, video indexing, and image processing.

SHUAI YANG received the B.S. degree in automation from the Changzhou Institute of Technology, Changzhou, China, in 2018. He is currently pursuing the M.S. degree in control theory and control engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and deep learning.

CHAOWEI LIN received the B.S. degree in mechanical design, manufacturing, and automation from the Guangdong University of Technology, Guangzhou, Guangdong, China, in 2018. He is pursuing the M.S. degree in control engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and deep learning.



HANQING CHEN received the B.S. degree in opto-electronics information science and engineering from the Changzhou Institute of Technology, Changzhou, China, in 2018. He is currently pursuing the M.S. degree in control engineering with the University of Shanghai for Science and Technology, Shanghai, China. His research interests include computer vision and deep learning.

KOJI KOTANI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Tohoku University, Japan, in 1988, 1990, and 1993, respectively, all in electronic engineering. He is currently a Professor with the Department of Intelligent Mechatronics, Akita Prefectural University. He is also involved in the research and development of high-performance devices/circuits, as well as intelligent electronic systems. He is a member of the Institute of Electronics, Information, and Communication Engineers of Japan.

QIU CHEN (Member, IEEE) received the Ph.D. degree in electronic engineering from Tohoku University, Japan, in 2004. Since then, he has been an Assistant Professor and an Associate Professor with Tohoku University. He is currently a Professor with Kogakuin University. His research interests include pattern recognition, computer vision, information retrieval, and their applications. He serves on the editorial boards of several journals, as well as committees for a number of international conferences.