CFRLA-Net: A Context-aware Feature Representation Learning Anchor-free Network for Pedestrian Detection

Jun Li, Yuquan Bi, Sumei Wang, and Qiming Li*,

***Abstract*—High resolution and strong semantic representation are both vital for feature extraction networks of pedestrian detection. The existing high-resolution network (HRNet) has presented a promising performance for pedestrian detection. However, we observed that it still has some significant shortcomings for heavily occluded and small-scale pedestrians. In this paper, we propose to address the shortcomings by extracting semantic and spatial context from HRNet. Specifically, we propose a Context-aware Feature Representation Learning Module (CFRL-Module), which combines a Multi-scale Feature Context Extraction Parallel Block for Convolution and Self-attention (CEPCA-Block) with two parallel paths and an Equivalent FFN (EFFN) Block. The core CEPCA-Block adopts a parallel design to integrate convolution and multi-head self-attention (MHSA) with low parameter computational cost, which can obtain the deep semantic context by convolution path and precise context by MHSA path. Furthermore, to overcome the inefficiency of global MHSA in high-resolution pedestrian detection, we propose a novel local window MHSA, which can significantly reduce memory consumption but barely affect the detection performance. Cascading the proposed CFRL-Module with the anchor-free detection head constitutes our Context-aware Feature Representation Learning Anchor-Free Network (CFRLA-Net). The proposed CFRLA-Net can catch a high-level understanding of the heavily occluded and small-scale pedestrian instances based on HRNet, which can effectively solve the limitation of the insufficient feature extraction ability of HRNet for the hard samples. Experimental results show that CFRLA-Net achieves state-of-the-art performance on CityPersons, Caltech, and CrowdHuman benchmarks.**

***Index Terms*—pedestrian detection, HRNet, context, self-attention, anchor-free, occluded and small-scale pedestrians.**

I. Introduction

PEDESTRIAN detection, as a particular branch of general object detection, has attracted increasing interest in computer vision and multimedia analysis communities [1–12]. It plays an important role in various practical applications, such as automotive driving systems [1–3], human-robot interaction [4, 5], and intelligent video surveillance [6–10]. Over the past few years, the performance of pedestrian detection has been greatly improved with the development of Convolutional Neural Network (CNN). However, the pedestrian detection task still remains a challenging problem because of the large variety of scales, the low resolution of small-size targets, and occlusion issues.

The mainstream pedestrian detection methods in recent years can be divided into anchor-based [20–29, 39–41, 45] and anchor-free [30, 31, 53, 67] two categories. The anchor-based detectors represented by Faster-RCNN are performed by classifying and regressing anchor boxes. Thus, the design of the hyper-parameters (*e.g.*, size, aspect ratio, and number) for the anchor box has a great impact on the detection performance. Compared to anchor-based detectors, anchor-free detectors directly detect pedestrians from input images without enumerating a large number of candidate proposals. In addition, without the handcrafted design with respect to the size and aspect ratio of anchors, the anchor-free detectors have a good generalization ability on different detection works. Whichever kind of method is adopted, most of the existing pedestrian detectors use CNN-based feature extraction networks to learn pedestrian features. However, different from image classification and object detection tasks, there are a large number of occluded and small-scale samples in pedestrian detection. The ability of networks for feature representation learning of these hard samples has a significant impact on the performance of pedestrian detectors.

There are mainly two common techniques to improve the networks′ ability to learn the complete pedestrian features. One is generating a high-resolution heatmap with high quality for accurate local discrimination of small-scale and occluded pedestrians. The other is extracting strong semantic information for full-scene visual understanding, which can ensure overall prediction accuracy. However, since stronger semantic information often means lower resolution, it seems difficult to achieve a good trade-off between high resolution and strong semantic information. The recently proposed high-resolution network (HRNet) [64] has presented a promising performance for pedestrian detection, because it can learn high-resolution representations and strong semantic information, while well ensuring the information transmission between low and high resolution feature maps. However, we observed that HRNet still has insufficient learning ability for difficult pedestrian instances, especially for heavily occluded and small-scale pedestrians. For example, in Fig. 1 (a), the center heatmap responses of heavily occluded pedestrian instances are weak

This work was supported in part by the National Natural Science Foundation of China under Grant 62102394, and in part by the Natural Science Foundation of Fujian Province of China under Grant 2020J05083, and in part by the Science and Technology Program of Quanzhou under Grant 2020C052.

Yuquan Bi is with the department of Advanced Manufacturing, Fuzhou University, Quanzhou, China (e-mail: yuquanbi@yeah.net). Jun Li and Qiming Li are with the Laboratory of Robotics and Intelligent Systems, Quanzhou Institute of Equipment Manufacturing, Haixi Institute, Chinese Academy of Sciences, Quanzhou, Fujian, 362216, China (e-mail: junli@fjirsm.ac.cn, qimingli@fjirsm.ac.cn). Sumei Wang is with the Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China (e-mail: may.sm.wang@polyu.edu.hk).
*Corresponding author.

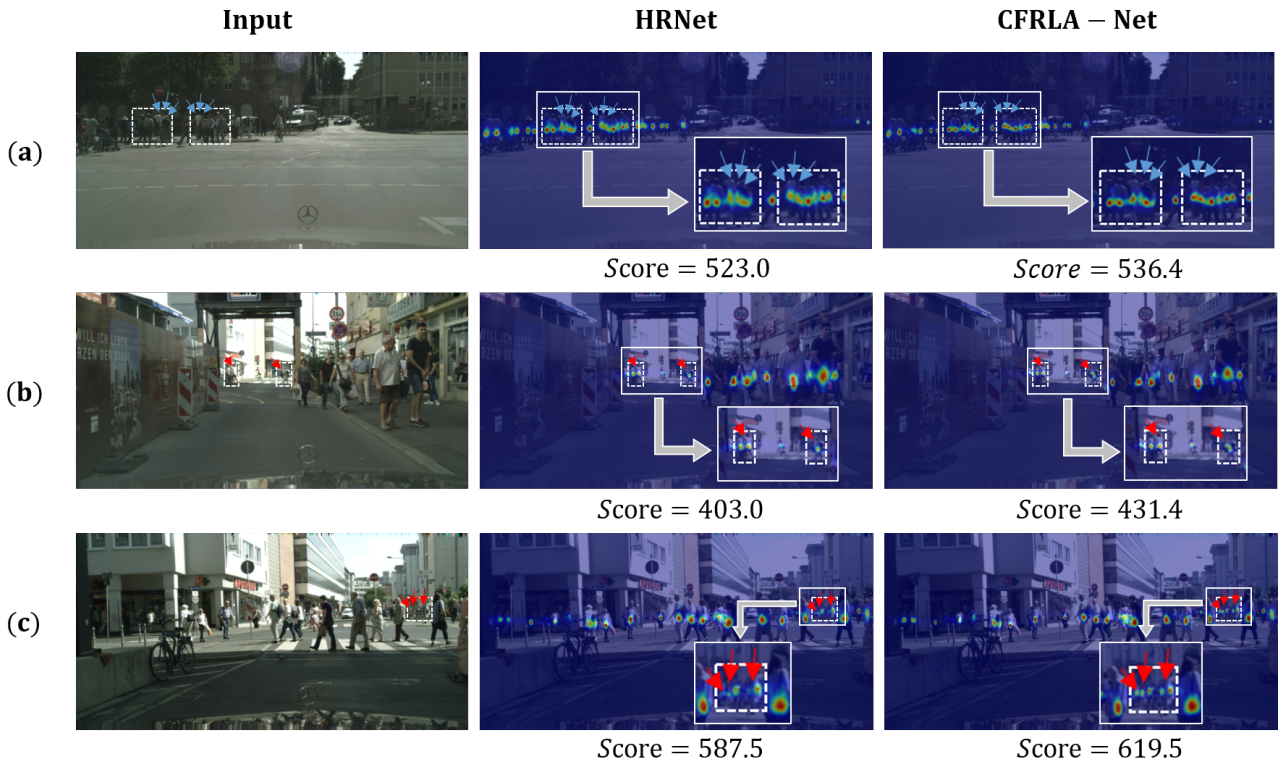

Fig. 1: **Illustration of the pedestrian detection center heatmap results of HRNet and CFRLA-Net.** We use white dashed boxes to mark areas where the results are significantly different. Compared with HRNet, our CFRLA-Net has stronger response strength and more precise response location to center heatmap of occluded pedestrian instances (Blue arrows). And our CFRLA-Net can detect small-scale pedestrian instances missed by HRNet (Red arrows). Furthermore, the heatmap score of our method outperforms the baseline, which proves that our method has stronger response strength.

(Blue arrows in the second column). While in (b) and (c), some small-scale pedestrian instances are missed (Red arrows in the second column). We suppose that it could be attributed to the macrostructure of HRNet, which can be summarized as two aspects below: (i) The four branches of HRNet are not deep enough, which limits the capacity of extracting strong semantic information. (ii) Some unaligned and useless spatial information is generated during the information transmission process, which is detrimental to the precise position of occluded and small-scale pedestrians. As proven by numerous computer vision literature [16–18], context is a statistical property of the world we live in and provides critical information to help us understand complex visual scenes (especially when objects are small and occluded in images) faster and more accurately. It has been proved in numerous pedestrian detection literature [39–41] that deep convolution is a natural way to obtain semantic context (represents the possibility of the existence of an object [16]), which can solve the insufficient ability of HRNet to obtain strong semantic information. Although few studies involve the application of spatial context (represents the position of an existing object [16]) in pedestrian detection, the success of spatial context in pedestrian re-identification [19] inspires us to explore its benefit in pedestrian detection. [19] points out that the spatial relationships between pedestrians in the whole scene image will contribute to more discriminative representations, which may become the key to solving the unaligned and useless spatial information in HRNet. Benefiting from the great success of the Transformer Networks in Computer Vision [59, 60], the core multi-head self-attention (MHSA) operation shows a powerful capability of obtaining precise spatial relationships. Thus, we desire to apply MHSA operation to pedestrian detection for spatial context extraction. Based on the above theories, we point out that more discriminative representation learning can be achieved by extracting semantic context and spatial context, which can effectively solve the two shortcomings of HRNet for occluded and small-scale pedestrians without destroying the macrostructure of the network.

With the motivation of designing an efficient module to obtain deep semantic context by convolution operation and precise spatial context by MHSA operation, we propose a simple but effective CFRL-Module. The newly proposed CFRL-Module consists of a CEPCA-Block and an EFFN Block, which is appended to each feature map of the feature pyramid formed by the HRNet stage4. Based on the theory of the ”Shift” [54–56] and some works [57, 58] to combine convolution and MHSA through a parallel mechanism, we design a CEPCA-Block that adopts a parallel design, which can integrate MHSA and convolution with low computational cost. Specifically, we first project the input feature maps with a 1×1 convolution and obtain the intermediate features shared by the convolution path and MHSA path. In this way, we can

get as many shared parameters as possible from the parallel paths, which can greatly reduce the parameter computational cost. Then, we design the two parallel paths according to their respective paradigms to better learn context information from HRNet. For the convolution path, we adopt a Group Convolution to obtain the deep semantic context. And for the MHSA path, we design a Multi-Resolution MHSA with relative-distance-aware position encoding proposed in [61, 62] to obtain the precise spatial context. Furthermore, to resolve the disadvantage that the Multi-Resolution MHSA consumes too much memory in high-resolution features of the feature pyramid, we propose a novel local window MHSA with a specific aspect ratio window, which can greatly reduce the memory consumption but barely affect the detection performance. After that, we get a strong semantic representation and provide information communications across channels for the output feature maps of CEPCA-Block through the EFFN Block. The transformed multi-scale feature maps by CFRL-Module are resized to the same resolution and concatenated as the input of the detection head. Finally, as in the typical anchor-free detectors, pedestrian detection is formulated as a center and scale prediction task in the detection head by conducting convolution on the concatenated feature maps. Cascading the proposed CFRL-Module with the anchor-free detection head constitutes our Context-aware Feature Representation Learning Anchor-free Network (CFRLA-Net), which can effectively alleviate the limitation of the insufficient feature extraction ability of HRNet for the heavily occluded and small-scale pedestrian instances. As shown in Fig. 1, our CFRLA-Net shows a great advantage than the baseline, which can intuitively demonstrate that two shortcomings of HRNet for occluded and small-scale pedestrians can be effectively alleviated by better learning the semantic and spatial context from HRNet. Specifically, our CFRLA-Net has stronger response strength and more precise response location than HRNet (Blue arrows in the third column). And our CFRLA-Net can detect small-scale pedestrian instances missed by HRNet (Red arrows in the third column). Furthermore, to reflect the response strength of the center heatmap, we define a variable called the heatmap score, which represents the sum of the heatmap confidences of the maximum response channel of the center heatmap. It can be seen that the heatmap score of our method outperforms the baseline, which proves that our method has stronger response strength.

In summary, the main contributions of this work are summarized as follows:

- Aiming at the poor detection performance of HRNet for heavily occluded and small-scale pedestrian instances, we propose a new CFRL-Module that adopts a parallel design to obtain the deep semantic context by convolution operation and precise spatial context by MHSA operation. We also investigate the impact of the two parallel paths on the detection performance separately.
- To greatly reduce the memory consumption introduced by the MHSA path in CFRL-Module, we propose a novel local window MHSA for high-resolution features with barely performance loss. We also investigate the influence of the novel local window MHSA with different aspect ratio windows on the representation ability of the human body.
- The proposed CFRLA-Net achieves state-of-the-art performance validated by extensive experiments on three challenging benchmarks (*i.e.*, CityPersons [14], Caltech [13], and CrowdHuman [15]).

II. Related Work

A. Generic Object Detection

With the development of CNN, object detection has obtained great progress recently. CNN-based object detectors can be broadly divided into two categories: anchor-based detectors [20, 27–29, 42–45] and anchor-free detectors [31, 46–53]. Anchor-based detectors first pre-define a set of anchors at different scales and then classify these candidate anchor boxes to find the object instances. Most existing anchor-based detectors can be classified into one-stage [20, 42, 45, 66] or two-stage [27–29, 43, 44] two categories. Faster R-CNN [44] is one of the most representative two-stage frameworks, which first generates a sparse set of regions of interest (RoI), and then refines them with classification and regression networks. On the contrary, one-stage anchor-based detectors (*e.g.*, SSD [42], RetinaNet [66], and ALFNet [45]) remove the RoI pooling step and directly detect objects in a single network.

Anchor-free detectors can bypass the requirement of anchor boxes and directly find and locate objects from the input images. As one of the most popular anchor-free detectors, YOLO [46] utilizes points around the center of an object to predict the bounding boxes. Afterwards, many key-points based anchor-free detectors [31, 47–53] are developed, where the goal is to predict key-points of the bounding box. For instance, CornerNet [48] models information of top-left and bottom-right corners with novel feature embedding methods and corner pooling layer to correctly match keypoints belonging to the same objects. CenterNet [50] first predicts bounding boxes by pairs of corners and then estimates center probabilities of the initial prediction to reject easy negatives. CSP [31] transforms the task of detection into the predictions of centers and their corresponding scales.

B. Pedestrian Detection

Pedestrian detection is one of the most important topics in computer vision since human is the central component in real-world applications. Many efforts [20–30, 32–38, 45] have been made to solve the difficulty of extracting complete pedestrian features caused by occluded and small-scale pedestrians. These works can be mainly categorized into optimization-based methods [30, 32–38] and part-based methods [20–30, 45]. Optimization-based methods handle the difficulty by designing some special losses or variants of Non-Maximum Suppression (NMS). The local parts of human body play an important role in representing discriminative features for occluded pedestrians, thus some newly proposed part-based methods build part detectors to parse complete pedestrian features.

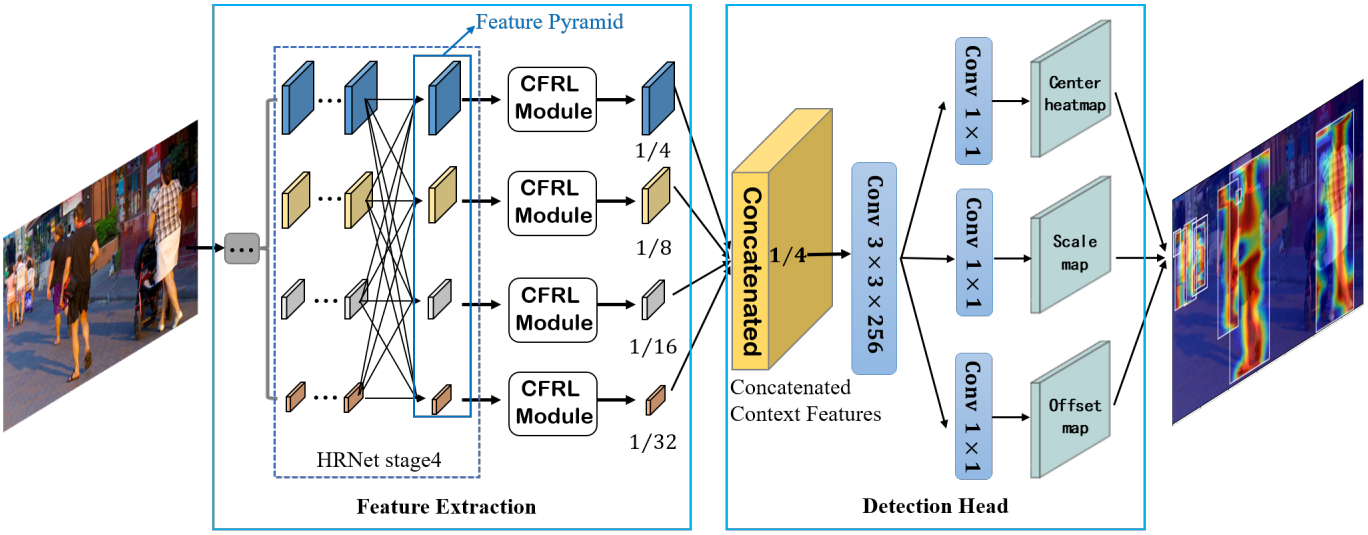

Fig. 2: **The overall architecture of our CFRLA-Net.**

Apart from the above two typical categories, an intuitive idea to learn complete pedestrian features is to use the context information. Therefore, some context-aware methods capturing useful context information and fusing them in multi-scale feature maps through carefully designed feature fusion strategies are proposed. Ren *et al.* [39] build a recurrent rolling convolution architecture to gradually aggregate context information from different layers. Zhang *et al.* [40] concatenate the RoI from multi-scale features along with the global context. Fei *et al.* [41] develop a new pixel-level context embedding module by integrating multi-cue contexts into a deep CNN feature hierarchy. However, these works focus on extracting semantic context for semantic guidance to enhance the features of pedestrians. They ignore the importance of precise spatial context for occluded and small-scale pedestrian detection. In our work, we adopt a parallel design to integrate convolution and MHSA to simultaneously obtain deep semantic context and precise spatial context, which can make the detector more robust to occluded and small-scale pedestrians.

C. Shift and Self-attention

"Shift" is a model compression operation proposed for deep neural network acceleration [54–56]. Its development provides theoretical support for the combination of convolution and self-attention. Wu *et al.* [54] prove that shift operation can be seen as a special case of depthwise convolution, which means spatial convolution can be alternated by spatial shift operation to shift feature maps and by pointwise convolution to aggregate spatial information. You *et al.* [56] adopt the theory that multiplication can be replaced by additions and logical bit-shifts. They yield a new type of deep network that involves only bit-shift and additive weight layers.

Based on the theory of the "Shift", the spatial convolution can be replaced by efficient shift and addition operations, which is similar to the self-attention paradigm. Thus, some researchers adopt a parallel design mechanism to combine self-attention and convolution instead of a simple serial design. For example, Pan *et al.* [58] prove that there exists a strong underlying relation between convolution and self-attention. Moreover, they deploy their parallel model on PVT [59], Swin Transformer [60], etc. Chen *et al.* [57] combine local-window self-attention with depth-wise convolution in a parallel design and propose bi-directional interactions across branches to provide complementary clues in the channel and spatial dimensions of both operations. However, both of them focus on complementing transformer models with convolution operations to introduce additional inductive biases, which means that structural deviations from input data are hardly assumed. This leads to the difficulty of training on small-scale datasets. Due to the insufficient data volume of the pedestrian dataset, it seems that the parallel design based on transformer frameworks is difficult to perform well for pedestrian detection. Unlike transformer frameworks, we integrate the MHSA module into the CNN-based framework by designing a CFRL-Module that adopts a parallel design mechanism to improve the ability of context learning.

III. CFRLA-Net

A. Review of HRNet

HRNet is first proposed in [64] for human pose estimation. [64] also proves that HRNet can work well on many other vision tasks. Recently, HRNet has presented a promising performance for pedestrian detection because it is strong not only at high-level semantic representation but also at low-level spatial detail. However, HRNet still has difficulty detecting a large number of occluded and small-scale pedestrian instances. To sum up, we find that HRNet has several drawbacks below for occluded and small-scale pedestrians: (i) The four branches of HRNet are not deep enough which limits the capacity of extracting semantic information. (ii) Some unaligned and useless spatial information is generated during the information transmission process, which is detrimental to the precise position of occluded and small-scale pedestrians.

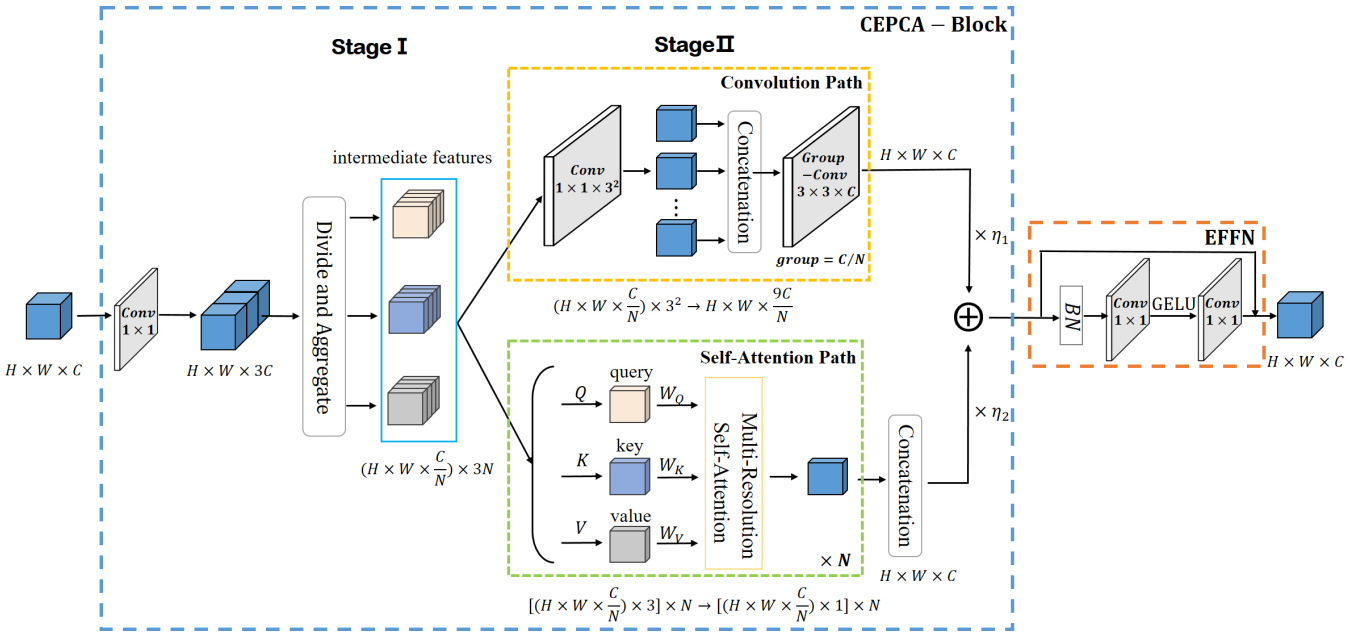

Fig. 3: **The architecture of our CFRL-Module.** Including CEPCA-Block and EFFN Block. The core CEPCA-Block consists of two stages. At Stage I, we obtain the intermediate features shared by the parallel structure. At Stage II, through parallel convolution path and MHSA path, we obtain context information of the intermediate features.

B. Architecture of CFRLA-Net

To alleviate the shortcomings of HRNet, we propose a simple but effective anchor-free network named CFRLA-Net, which can better learn the context information to make detectors more robust to occluded and small-scale pedestrians. The overall architecture of the CFRLA-Net is illustrated in Fig. 2. We will describe the details of CFRLA-Net from the following four perspectives: CFRL-Module, Feature Fusion, Pedestrian Dection Head, and Joint Optimization and Inference.

C. CFRL-Module

The objective of the CFRL-Module is to learn the deep semantic context and the precise spatial context from HRNet. Our CFRL-Module is composed of two key parts: a new CEPCA-Block that adopts a parallel design, which can integrate the convolution path and MHSA path with low computational cost; an EFFN Block is designed to get a strong semantic representation and provide information communications across channels for the output feature maps of the CEPCA-Block. The architecture of CFRL-Module is illustrated in Fig. 3.

1) CEPCA-Block

To obtain deep semantic context and precise spatial context, we propose a new CEPCA-Block that adopts a parallel design to integrate the convolution path and MHSA path with low computational cost. Furthermore, we also explore the impact of the two parallel paths on detection performance separately in Sec. IV-B. Although [58] also employed two paralleled paths based on self-Attention and convolution, there are several significant differences between us: (i) In order to obtain the deep semantic context, we deepen the channel and extract the strong semantic context in the convolution path, while [58] use identity map instead of deepening the channel. (ii) In order to obtain the precise spatial context, we adopt the relative-distance-aware position encoding proposed in [61, 62], which is not involved in [58]. (iii) Unlike various downstream vision tasks mentioned in [58], pedestrian detection task is highly dependent on high-resolution feature input. Thus, in order to solve the disadvantage of excessive memory consumption of high-resolution features, we propose a novel local window MHSA conforming to the characteristics of pedestrian feature distribution, which can greatly reduce memory consumption but barely affect detection performance. Therefore, our CEPCA-Block is well-designed for better learning the context of occluded and small-scale pedestrians based on HRNet.

As illustrated in Fig. 3, we divide the process of CEPCA-Block into two stages. At stage I, we first adopt a 1×1 convolution to expand the input feature maps to $3 \times C$ feature maps to get the deep semantic information. Then, we divide the $3 \times C$ feature maps into N pieces and aggregate a rich set of intermediate features containing $\left(H \times W \times \frac{C}{N}\right) \times 3N$ feature maps, which are shared by the parallel structure in stage II. At stage II, convolution and MHSA lie in two parallel paths. Both of the paths take the intermediate features as the input and reorganize these intermediate features to meet their respective paradigms. For the convolution path with kernel size 3, we adopt a 1×1 convolution layer to change the $3N$ intermediate features to 3^2 projected feature maps. Following the shift and addition operations proposed in [56], we can process the projected feature maps in a convolution manner and gather information from a local receptive field like the traditional ones. However, the mentioned operations are difficult to achieve vectorized implementation, which greatly impairs the actual efficiency. Thus, we take the group convolution with learnable kernel weights proposed in [58] to replace the inefficient shift and addition operations. Specifically, we first

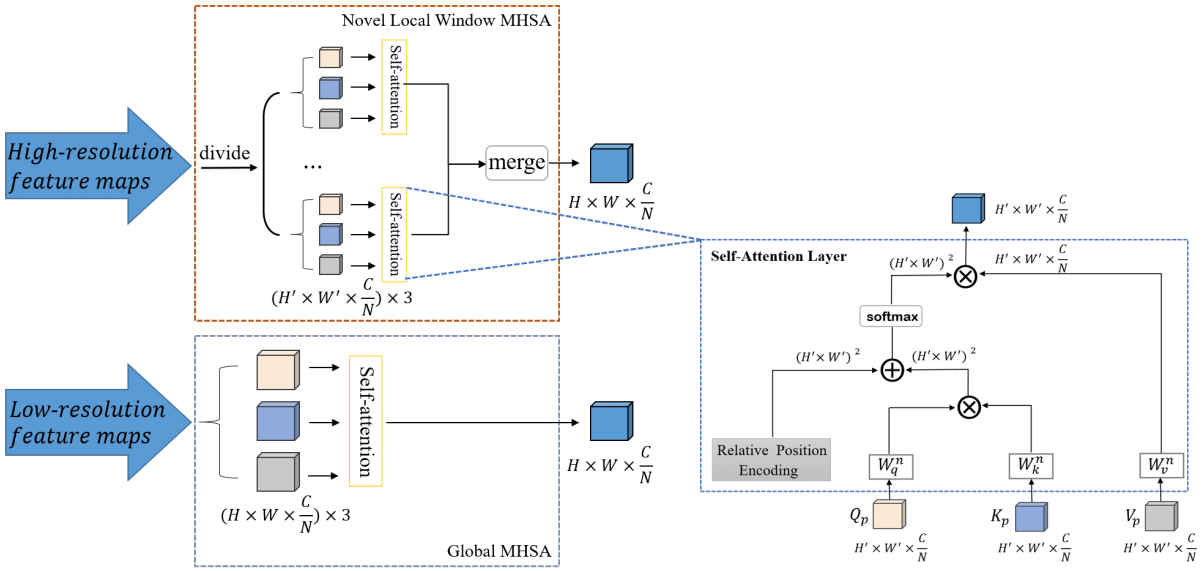

Fig. 4: **Illustration of the Multi-Resolution Self-Attention in CFRL-Module.** Including Novel Local Window MHSA for high-resolution feature maps and Global MHSA for low-resolution feature maps. In Novel Local Window MHSA, we use local window with a specific aspect ratio of 1:2 and relative-distance-aware position encodings.

concatenate the projected feature maps to get $9\times C\backslash N$ channel feature map. Then, we release the group convolution kernel as learnable weights, with the carefully designed kernel weights proposed in [58] as initialization. Finally, we apply the already initialized group convolution with kernel size 3, C filters, and $C\backslash N$ group to get C channel output feature map F_{conv}, which can obtain the deep semantic context. For the MHSA path, we gather the intermediate features into N groups, where each group contains three pieces of features. Afterwards, we take the three pieces of features as queries $Q \in \mathbb{R}^{H\times W\times \frac{C}{N}}$, keys $K \in \mathbb{R}^{H\times W\times \frac{C}{N}}$, and values $V \in \mathbb{R}^{H\times W\times \frac{C}{N}}$. We further take the N groups as the number of heads. According to the new designed Multi-Resolution Self-Attention Module composed of the global MHSA or the novel local window MHSA (Shown in Fig. 4) discussed below, we get $C\backslash N$ channel feature maps of the MHSA and concatenate all of the feature maps to get C channel output feature map F_{att}, which can obtain the precise spatial context. Finally, we add the output feature maps from both paths with two dynamic weights η_1 and η_2, which are learnable and adjusted by backpropagation during training:

$$F_{out} = \eta_1 F_{conv} + \eta_2 F_{att}. \tag{1}$$

Note that η_1 and η_2 practically reflect the model's bias towards convolution or MHSA at different features. We will further discuss the influence of these two parameters in Sec. IV-B. Through CEPCA-Block, we can obtain the deep semantic and precise spatial context of HRNet output features, which can make up for the lack of HRNet's ability to extract occluded and small-scale pedestrian features.

2) Global MHSA

To efficiently aggregate the context contained in low-resolution feature maps at stage4 of HRNet-32 with the down-sampling rates of 16 and 32, we adopt global MHSA that has a global receptive field and powerful context capture capability. As illustrated in the bottom row of Fig. 4, we perform MHSA on N heads Q, K, V features:

$$\text{Multihead}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{Concat}\left[\text{head}_1,\ldots,\ \text{head}\ _N\right]\mathrm{W}_0, \tag{2}$$

where $\text{head}_n \in \mathbb{R}^{H\times W\times \frac{C}{N}}$ can be defined as:

$$\text{head}_n = \text{softmax}\left[\frac{\left(\mathrm{Q}W_q^n\right)\left(\mathrm{K}W_k^n\right)^T}{\sqrt{C/N}} + B\right]\mathrm{V}W_v^n, \tag{3}$$

where the projection matrices for queries, keys and values are $W_q^n \in \mathbb{R}^{\frac{C}{N}\times C}$, $W_k^n \in \mathbb{R}^{\frac{C}{N}\times C}$, and $W_v^n \in \mathbb{R}^{\frac{C}{N}\times C}$ for $\mathrm{n} \in \{1,\ldots,\mathrm{N}\}$; C represents the number of channels; $W_0^n \in \mathbb{R}^{C\times C}$ is the multi-head weight matrix; H represents the height of input, and W represents the width of input. B represents the relative-distance-aware position encoding proposed in [61, 62], which can effectively associate information across objects with positional awareness. Finally, we can obtain the output out_{glo}^{MHSA} of the global MHSA:

$$\text{out}_{\text{glo}}^{\text{MHSA}} = \text{Multihead}\ (\mathrm{Q},\ \mathrm{K},\ \mathrm{V})\,. \tag{4}$$

3) Novel Local Window MHSA

Although global MHSA has a global receptive field and powerful context capture capability, the memory consumption is catastrophic in the high-resolution pedestrian detection task. Thus, for high-resolution feature maps at stage4 of HRNet-W32 with the down-sampling rates of 4 and 8, we propose a novel local window MHSA to replace the global MHSA, which can greatly reduce the memory consumption with barely performance loss. The window size of the novel local window MHSA is designed with a specific aspect ratio of 1:2, which has been proved in our experiments that it can achieve high performance with low memory consumption, the setting of the window size will be discussed in the ablation studies.

Specifically, as illustrated in the top row of Fig. 4, we

first divide the mentioned $Q, K, V \in \mathbb{R}^{H \times W \times \frac{C}{N}}$ into a set of non-overlapping local windows: $Q \rightarrow \{Q_1, Q_2, \ldots, Q_p\} \in \mathbb{R}^{H' \times W' \times \frac{C}{N}}$, where H' and W' represent the height and width of the local window. K and V are same as Q. Different from the traditional local window self-attention that has an identical window size, we set the window to an aspect ratio of 1:2, which is suitable for peculiarities in the aspect ratio of pedestrians to represent the distributions of pedestrian features.

Then, we perform MHSA on N heads Q, K, V features within each window independently. The formulation of MHSA on the $p-th$ window is given as:

$$\text{Multihead}\left(\mathrm{Q}_p, \mathrm{K}_p, \mathrm{V}_p\right) = \text{Concat}\left[\text{head}_1^p, \ldots, \text{ head }_N^p\right] \mathrm{W}_0, \tag{5}$$

as illustrated in the right column of Fig. 4, $\text{head}_n^p \in \mathbb{R}^{H' \times W' \times \frac{C}{N}}$ can be defined as:

$$\text{head}_n^p = \text{softmax}\left[\frac{\left(\mathrm{Q}_p W_q^n\right)\left(\mathrm{K}_p W_k^n\right)^T}{\sqrt{C/N}} + B\right] \mathrm{V}_p W_v^n, \tag{6}$$

where C, W_0^n, W_q^n, W_k^n, W_v^n and B are same as the corresponding parameters in the global MHSA. Through relative-distance-aware position encoding, we can get the relative distances between features at different locations so that it can incorporate the relative position information into the local window MHSA to obtain the precise spatial context.

With MHSA aggregates information in each window, we merge all of the non-overlapping local windows to obtain the output out_{loc}^{MHSA} of the novel local window MHSA:

$$\text{out}_{\text{loc}}^{\text{MHSA}} = \|_{l=1}^{p}\left(\text{ Multihead }\left(\mathrm{Q}_l,\ \mathrm{K}_l,\ \mathrm{V}_l\right)\right), \tag{7}$$

where $\|$ is the mergence of the outputs of p non-overlapping local windows.

4) EFFN Block

MHSA performs self-attention separately within non-overlapping channels, thus there are no information communication across channels. To get strong semantic representation and provide information communications across channels, we desire to use 1×1 layers to increase the depth. Following the FFN proposed in the Transformer network [57, 59], we design an EFFN Block composed of shortcut, BN, two 1×1 convolutions, and GELU. As a common practice, the number of internal channels of EFFN Block is $4\times$ as the input. We place the EFFN Block after the CEPCA-Block and take the output of EFFN BLock as the final output of the CFRL-Module.

D. Feature Fusion

As the combination of deeper feature maps is helpful to improve detection performance [31], we feed the context features formed by CFRL-Module into the deconvolution layer to resize them to the same resolution (*i.e.* $W/r \times H/r$, where input image $I \in \mathbb{R}^{W \times H \times 3}$ and $r = 4$ is the down-sampling rate that performs best as proved in [31]). Afterwards, these resized feature maps are concatenated and then fed into the pedestrian detection head. It is worth to point out that more complicated feature fusion strategies like [39–41] may further improve the detection performance, but it is not in the scope of our paper.

E. Pedestrian Detection Head

In the detection head, we parse the obtained concatenated feature maps into detection results. Specifically, a 3×3 convolutional layer is firstly attached to reduce the channel dimension of the concatenated feature maps to 256, and three sibling 1×1 convolution layers are utilized to generate the center heatmap, scale map, and offset map, respectively. Afterwards, according to the center heatmap and corresponding scales in the scale map, the bounding boxes of pedestrians of input images can be generated automatically. Finally, the detection performance is further improved via the offset prediction branch that can adjust the center locations of pedestrians slightly.

We also need to construct the ground truth for each output map in the detection head. Follow the bounding box annotations of input images given in [31], we can generate the center, scale, and offset ground truth automatically.

1) Center Ground Truth

For the center ground truth, the location where a pedestrian's center point falls is assigned as positive. However, it is difficult to decide an exact center point of pedestrian for training. Thus, we apply a 2D Gaussian distribution $P(\cdot)$ centered at the location of each positive, which is formulated as:

$$\begin{aligned} M_{ij} &= \max_{n=1,2,\ldots,K} P\left(i, j; x_n, y_n, \sigma_{w_n}, \sigma_{h_n}\right) \\ P\left(i, j; x, y, \sigma_w, \sigma_h\right) &= e^{-\left(\frac{(i-x)^2}{2\sigma_w^2} + \frac{(j-y)^2}{2\sigma_h^2}\right)}, \end{aligned} \tag{8}$$

where (i, j) denotes the location of the output center heatmap of our network; K denotes the number of pedestrians in the input image; σ_w and σ_h denote the variances of 2D Gaussian distribution, which are proportional to the width and height of pedestrians respectively; (x_n, y_n, w_n, h_n) denotes the center coordinates, width, and height of the $n-th$ pedestrian.

2) Scale Ground Truth

The scale ground truth can be generally defined as the height and/or width of human bodies. Line annotation is first proposed in [14], and can be used to generate the bounding box with a uniform aspect ratio of 0.41. According to the line annotation, our network predicts the height of each pedestrian and generates the bounding box with the predetermined aspect ratio (*i.e.*, 0.41). The scale ground truth can be defined as:

$$S_{ij} = \log\left(h_n\right). \tag{9}$$

To reduce the ambiguity, the negatives within a radius 2 of the centers of pedestrians are also assigned with $\log\left(h_n\right)$.

3) Offset Ground Truth

We follow [45, 48] to append an offset branch to adjust center locations accordingly before remapping. The offset ground truth can be formulated as:

$$O_{ij} = \left(\frac{x_n}{r} - \left\lfloor \frac{x_n}{r} \right\rfloor, \frac{y_n}{r} - \left\lfloor \frac{y_n}{r} \right\rfloor\right). \tag{10}$$

F. Joint Optimization and Inference

1) Optimization

For center predictions, to reduce the extreme positive-negative imbalance issue, we train the target L_{center} of center location prediction in the detection head as a classification task via the Focal Loss [66]:

$$L_{\text{center}} = -\frac{1}{N}\sum_{i=1}^{W/r}\sum_{j=1}^{H/r}\alpha_{ij}\left(1-\hat{p}_{ij}\right)^{\gamma}\log\left(\hat{p}_{ij}\right), \tag{11}$$

where $\hat{p}_{ij}$ and α_{ij} can be defined as

$$\begin{aligned}\hat{p}_{ij} &= \begin{cases} p_{ij} & \text{if } y_{ij}=1 \\ 1-p_{ij} & \text{otherwise} \end{cases} \\ \alpha_{ij} &= \begin{cases} 1 & \text{if } y_{ij}=1 \\ (1-M_{ij})^{\beta} & \text{otherwise,} \end{cases}\end{aligned} \tag{12}$$

where $y_{ij}=1$ denotes that a pedestrian′s center falls the location (i,j). Let $p_{ij}\in[0,1]$ denote the network′s predicted distribution indicating whether the location (i,j) is the center of a pedestrian or not. As in [31], to combat the positive-negative imbalance issue, the Gaussian mask M_{ij} is applied to reduce the contribution of negative samples to the full training loss. γ and β are the focusing hyper-parameters, which are set to 2 and 4 as suggested in [48, 66].

For scale and offset predictions, we train the targets L_{scale} and L_{offset} as a regression task via the Smooth L1 loss [44]:

$$\begin{aligned}L_{\text{scale}} &= -\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\text{Smooth}\,L1\left(s_n,\bar{s}_n\right) \\ L_{\text{offset}} &= -\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\text{Smooth}\,L1\left(o_n,\bar{o}_n\right),\end{aligned} \tag{13}$$

where s_n and $\bar{s}_n$ represent our network′s prediction and the ground truth of $n-th$ pedestrian′s scale; o_n and $\bar{o}_n$ represent our network′s prediction and the ground truth of $n-th$ pedestrian′s offset.

Finally, the full training loss is defined as:

$$L=\lambda_{\text{c}}L_{\text{center}}+\lambda_{\text{s}}L_{\text{scale}}+\lambda_{\text{o}}L_{\text{offset}}, \tag{14}$$

where λ_{c}, λ_{s}, and λ_{o} are the weights for the full loss function, which are experimentally set to 0.01, 1 and 0.1, respectively. When we minimize the loss, according to the back propagation algorithm, along with other parameters of the network, the parameters η_1, η_2 in Eq. 1 and the parameters W_q^n, W_k^n, and W_v^n of the novel local window MHSA and the global MHSA in Eq. 3 and Eq. 6 can be also optimized simultaneously.

2) Inference

The inference is as simple as forwarding an image through the backbone network. We first use a confidence threshold of 0.01 to filter out the locations with low confidence in the center heatmap, along with their corresponding scales in the scale map. Then bounding boxes are generated automatically with pedestrian′s centers and scales and are then adjusted according to the offset map before remapping to the original image size.

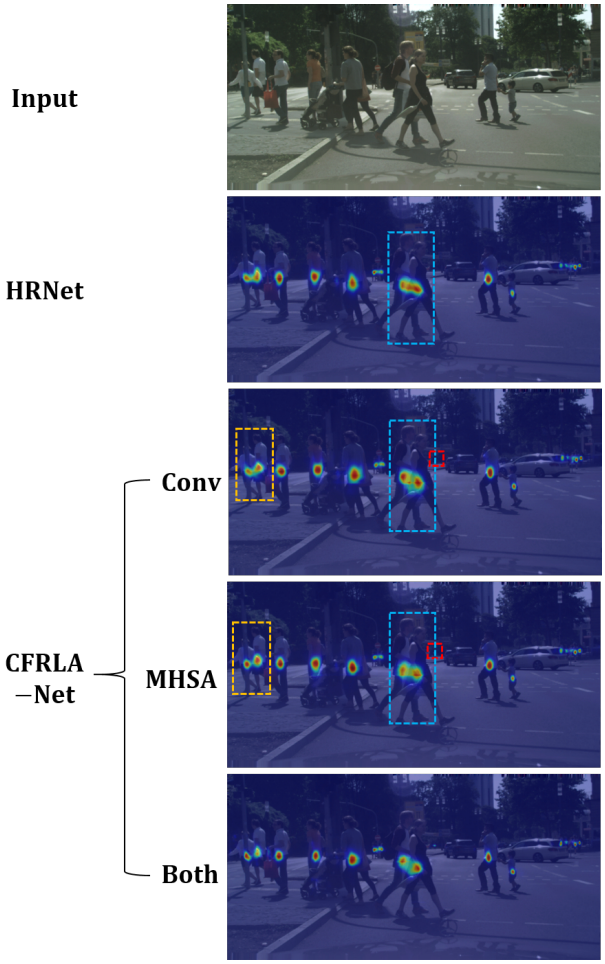

Fig. 5: **Illustration of the pedestrian detection center heatmap results of different paths in CEPCA-Block.**

Finally, NMS with a threshold of 0.5 is applied to yield the final pedestrian detection results.

IV. Experiments

A. Experimental Settings

1) Datasets and Evaluation Metrics

To fully validate the robustness of the proposed CFRLA-Net, we perform extensive experiments on three challenging pedestrian datasets: CityPersons [14], Caltech [13], and CrowdHuman [15]. CityPersons is a recently released challenging pedestrian detection dataset, which contains 5,050 images (2,975 for training, 500 for validation, and 1,575 for testing). As one of the predominant and representative pedestrian datasets, Caltech comprises approximately 2.5 hours of auto-driving video, which is divided into 42,782 training images and 4,024 testing images. We perform all the experiments on Caltech using the refined annotations provided in [68]. CrowdHuman is recently presented to specifically target crowd scenes, which collects 15,000, 4,370, and 5,000 images from the Internet for training, validation, and testing subsets respectively. There are $470k$ pedestrian instances in the training and validation subsets, and approximately 22.6

TABLE I: The MR^{-2} under four setups for our detector with separate path, $i.e.$, convolution path and MHSA path, in CEPCA-Block in comparison with the baseline on CityPersons (Red and blue indicate the best and second-best results).

Method	$MR^{-2}(\%)$					
	Reasonable	Heavy	Partial	Bare	GFLOPs	Params
CSP+HRNet-W32	10.4	48.1	9.3	6.8	135.8	29.4M
CFRLA-Net-Conv	9.4	44.6	8.2	5.2	137.3	29.9M
CFRLA-Net-Att	7.7	40.0	6.2	4.2	137.8	29.9M
CFRLA-Net	7.3	39.0	5.2	4.0	138.2	29.9M

TABLE II: The MR^{-2} under three setups for our detector with CFRL-Module in different multi-scale features in comparison with the baseline on CityPersons.

Method	$MR^{-2}(\%)$								
	Reasonable	Heavy	Partial	Bare	Small	Medium	Large	η_1	η_2
CFRLA-Net-Low	7.67	40.72	5.99	4.22	7.18	7.41	4.39	0.5831	0.3942
CFRLA-Net-High	7.89	41.64	6.70	4.42	5.32	6.58	5.19	0.4016	0.5724
CFRLA-Net	7.26	38.94	5.20	3.99	5.16	6.40	4.03	0.5014	0.4848

TABLE III: The MR^{-2} under five setups for our detector with different settings of the window size in the novel local-window self-attention in CEPCA-Block on CityPersons.

Exp Id	Scale	Window Size	$MR^{-2}(\%)$							
			Reasonable	Heavy	Partial	Bare	Small	Medium	Large	GFLOPs
1	1 : 1	20×20	7.81	41.00	6.84	4.29	6.43	7.11	4.34	137.9
2	1 : 1	40×40	7.25	39.16	5.66	3.73	5.65	7.38	3.96	139.1
3	1 : 2	20×40	7.26	38.94	5.20	3.99	5.16	6.40	4.03	138.2
4	2 : 1	40×20	7.29	39.42	6.03	4.00	6.13	6.48	4.10	138.2
5	1 : 4	20×80	7.22	38.87	5.78	3.64	5.30	6.43	4.11	139.1

TABLE IV: Performance of CFRLA-Net in comparison with some state-of-the-art methods on CityPersons.

Method	Backbone	$MR^{-2}(\%)$						
		Reasonable	Heavy	Partial	Bare	Small	Large	Test time
RepLoss [32]	ResNet-50	13.2	56.9	16.8	7.6	N.A.	N.A.	N.A.
TLL [73]	ResNet-50	15.5	53.6	17.2	10.0	N.A.	N.A.	N.A.
TLL+MRF [73]	ResNet-50	14.4	52.0	15.9	9.2	N.A.	N.A.	N.A.
FRCN+A+DT [24]	VGG-16	11.1	44.3	11.2	6.9	N.A.	N.A.	N.A.
OR-CNN [33]	VGG-16	12.8	55.7	15.3	6.7	N.A.	N.A.	N.A.
ALFNet [45]	ResNet-50	12.0	51.9	11.4	8.4	19.0	6.6	0.27s/img
FRCNN [14]	VGG-16	15.4	N.A.	N.A.	N.A.	25.6	7.9	N.A.
FRCNN+Seg [14]	VGG-16	14.8	N.A.	N.A.	N.A.	22.6	8.0	N.A.
WIDEERPERSON [71]	VGG-16	11.1	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
PedHunter [77]	ResNet-50	8.3	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
CSP [31]	ResNet-50	11.0	49.3	10.4	7.3	16.0	6.5	0.33s/img
CSP [31]	HRNet-W32	10.4	48.1	9.3	6.8	13.2	5.9	0.28s/img
NOH-NMS [35]	ResNet-50	10.8	53.0	11.2	6.6	N.A.	N.A.	0.28s/img
APD [67]	DLA-34	8.8	46.6	8.3	5.8	N.A.	N.A.	0.16s/img
BGCNet [74]	HRNet-W32	8.8	43.9	8.0	6.1	N.A.	N.A.	0.16s/img
PRNet [75]	ResNet-50	10.8	53.3	10.0	6.8	N.A.	N.A.	0.22s/img
DAGN [70]	ResNet-50	11.9	43.9	12.1	7.6	18.7	5.9	0.22s/img
CFRLA-Net (ours)	HRNet-W32	7.3	39.0	5.2	4.0	5.2	4.0	0.14s/img

persons per image, which is of much higher crowdedness compared with the previous datasets.

Following the common practice, the log Miss Rate (MR) averaged over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ is exploited as the evaluation metric for all three datasets. In addition, Recall and Average Precision (AP) are also included to evaluate the detection performance for CrowdHuman.

2) Training Details

The proposed CFRLA-Net is implemented in PyTorch and trained on 4 NVIDIA PCIe A100 GPUs. HRNet-W32 [64] pre-trained on ImageNet [63] is selected as our backbone network, and AdamW [69] solver is applied to optimize the network. The input sizes of the training images are set to 640×1280, 480×640, and 800×1200 for CityPersons, Caltech, and CrowdHuman, respectively. The window sizes of the novel local-window MHSA are set to 20×40, 20×40, and 25×50 for CityPersons, Caltech, and CrowdHuman, respectively. We use brightness variation, horizontal flip, cropping, and random scaling (between 0.5 to 1.5) as the data augmentation. For CityPersons and Caltech, our CFRLA-Net is trained for $300k$ iterations with a mini-batch of 4 images per GPU, and the initial learning rate starts at 2×10^{-4} and decays to 1×10^{-4} after $200k$ iterations. For CrowdHuman, our detector is trained for $200k$ iterations with an initial learning rate of 2×10^{-4}, which then decreases to 1×10^{-4} after $100k$ iterations.

B. Ablation Studies

In this section, to verify the effectiveness of our method, we conduct the ablation studies of our detector on CityPersons. We first explore the influence of the convolution path and the MHSA path in the CFRL-Module on the detection results. Then, we demonstrate the impact of the proposed CFRL-Module using different resolution features in the feature pyramid on the performance improvement of our detector respectively. Finally, we conduct the ablative analysis of the window size setting of the novel local window MHSA in CEPCA-Block and fix it to its best-fit value in all experiments.

1) Impact of Different Parallel Paths in CEPCA-Block

To explore the effect of the convolution path and the MHSA path on the detection performance in the CEPCA-Block with parallel design, we remove the convolution path (denoted as CFRLA-Net-Att) and MHSA path (denoted as CFRLA-Net-Conv) respectively. As can be seen in Table I, the performance of CFRLA-Net-Att is close to that of CFRLA-Net, which proves that the MHSA path plays a dominant role in our CEPCA-Block. This proves that extracting the precise spatial context through the MHSA path is more beneficial for occluded and small-scale pedestrian instances than extracting the deep semantic context through the convolution path. Furthermore, the performance of CFRLA-Net-Conv is higher than CSP+HRNet-W32 on Heavy, Partial, and Bare subsets, which proves that the deep semantic context also brings certain gains to detector performance. In addition, CFRLA-Net-Conv, CFRLA-Net-Att, and CFRLA-Net have almost the same parameters overhead, which proves that convolution and MHSA can be integrated with low parameter computational cost through our careful design. To show the effect of network training more intuitively, we visualized the center heatmap of the above models in Fig. 5. As we can see, compared with HRNet, CFRLA-Net-Conv has a stronger and more precise heatmap response for large-scale pedestrians (blue dashed box). CFRLA-Net-Att can detect some small-scale pedestrians missed by CFRLA-Net-Conv (red dashed box) and provide a stronger response (orange dashed box). Due to the use of the multi-head and local windows, CFRLA-Net-Att has some discontinuous heatmap areas (blue dashed box). The CFRLA-Net can combine the advantages of the two paths and achieve a great performance improvement than CSP+HRNet-W32 with only 0.5M parameters higher.

2) Impact of CFRL-Module using different resolution features

To figure out the impact of the proposed CFRL-Module using different resolution features in the feature pyramid on the performance improvement of our detector respectively, we remove the novel local window MHSA CFRL-Module after the high-resolution features (denoted as CFRLA-Net-Low) in feature pyramid and the global MHSA CRFL-Module after the low-resolution features (denoted as CRFLA-Net-High) in feature pyramid, respectively. The fusion version of CFRLA-Net-Low and CFRLA-Net-High is our CFRLA-Net. The performance comparison of the above detectors is given in Table II. Since low-resolution feature maps correspond to the high-level semantic representation, CFRLA-Net-Low mainly focuses on semantic context extraction. And since high-resolution feature maps mainly correspond to low-level spatial detail, CFRLA-Net-High mainly focuses on spatial context extraction. To confirm our conclusion, we calculate the average of the parameters η_1 and η_2 in Eq. 1 in the test set. The result shows that $\eta_1 = 0.5831$ of the convolution path is higher than $\eta_2 = 0.3942$ of the MHSA path in CFRLA-Net-Low, which proves that the convolution path plays a more important role. And the result in CFRLA-Net-High is just the opposite. Thus, CFRLA-Net-Low with the deeper semantic context performs better than CFRLA-Net-High on Heavy, Partial, and Bare subsets. And CFRLA-Net-High with the more precise spatial context performs better than CFRLA-Net-Low on Small and Medium subsets. Due to the elegant fusion of CFRLA-Net-Low and CFRLA-Net-High, our CFRLA-Net achieves excellent performance on almost all subsets.

3) Parameter Analysis of Window Size in CEPCA-Block

To evaluate the influence of the window scale $H' : W'$ of the novel local window MHSA in CEPCA-Block on the detection performance, we test our detector trained with different scale settings of $H' : W'$ on CityPersons. It is worth to point that a larger window size will theoretically bring a larger receptive field and stronger context capture capabilities, but it will also bring exponentially increased memory consumption, thus our designed novel local window MHSA is aimed to greatly reduce memory consumption with barely performance loss. To verify the effectiveness of our design, we set up five experiments with different scale windows, denoted as experiment 1, ... , and experiment 5. Table III gives the performance comparison of our CFRLA-Net with different scale settings of window size in CEPCA-Block. As can be seen, with the increase of window

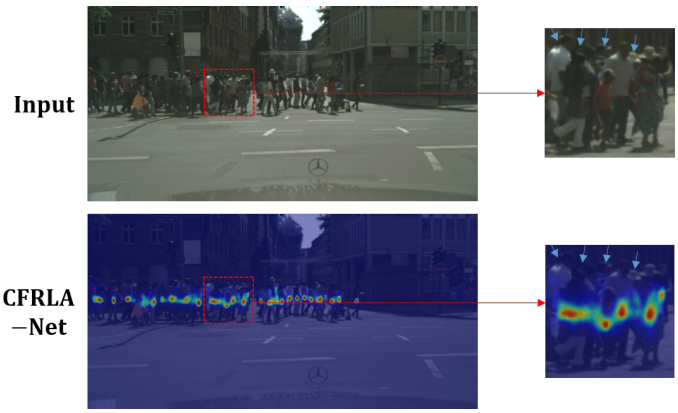

Fig. 6: **Illustration of the limitation of our detector.** Blue arrows represent pedestrian instances missed by our detector.

area, the performance on the Reasonable set will be better, thus experiments 2, 5 with the largest window area work better on the Reasonable set. However, both the settings cause excessive memory consumption and do not match the characteristics of pedestrian features. Our design of $H' : W' = 1 : 2$ can efficiently parse the global information of the human body that has a particular aspect ratio with much lower memory consumption (experiment 3), thus achieves excellent performance on Heavy, Partial, Bare, and Small subsets. As can be seen, our designed 20×40 window shows outperforming performance than the window area with the same modeling capabilities on all subsets (comparison between experiment 3 and experiment 4) and basically equal performance than the window area with higher modeling capabilities (comparison between experiment 3 and experiments 2, 5) with 0.9 GFLOPs lower. Therefore, to achieve an efficient balance between performance and memory consumption, we fix the window size scale $H' : W'$ of the novel local window self-attention in CEPCA-Block to be $1 : 2$.

C. Benchmark Comparison

1) CityPersons

Table IV shows the comparisons of our CFRLA-Net with some previous state-of-the-art pedestrian detectors [14, 24, 31–33, 35, 45, 67, 71, 73–75, 77]. It can be observed that CFRLA-Net beats the competitors and performs fairly well on all subsets. On the Reasonable subset, our CFRLA-Net achieves the best performance, with a gain of 1.0% MR^{-2} upon the closest competitor (PedHunter [77]). On the three occlusion subsets, our CFRLA-Net delivers the best performance among all the competing detectors even without any specific occlusion-handling strategies. Compared to some methods using occlusion handling strategies (*i.e.*, RepLoss [32], FRCN+A+DT [24], OR-CNN [33], NOH-NMS [35], and APD [67]), our method also provides superior detection performance. Compared with the closest competitor APD using an additional NMS strategy, our detector achieves a large performance gain of 7.6%, 3.1%, and 1.8% MR^{-2} respectively on three occlusion subsets. Moreover, for fair comparison with the baseline (*i.e.*, CSP+HRNet-W32), we also re-implement CSP with HRNet-W32. Our detector outperforms the baseline (CSP+HRNet-W32) consistently on almost all subsets and significantly improves Heavy, Partial, and Bare subsets by 9.1%, 4.1%, and 2.8% MR^{-2} respectively. It is worth to point that although our detector is not additionally designed for multi-scale pedestrians, it performs 10.6% MR^{-2} better than the closest competitor on the Small subset (16% of CSP) using only the most basic feature fusion strategy. These results illustrate that our detector has excellent performance for occluded and small-size pedestrians by better extracting the context information from HRNet and further prove that our method can efficiently solve the limitations of HRNet.

TABLE V: MR^{-2} performance of CFRLA-Net in comparison with some state-of-the-art methods on Caltech (PO-Partial Occlusion subset, HO-Heavy Occlusion subset).

Method	$MR^{-2}(\%)$		
	All	PO	HO
FasterRCNN [44]	62.6	8.7	53.1
HyperLearner [80]	61.5	5.5	48.7
RPN+BF [28]	59.9	7.3	54.6
ALFNet [45]	59.1	6.1	51.0
RepLoss [32]	59.0	5.0	47.9
OR-CNN+City [33]	58.8	4.1	45.0
RepLoss+City [32]	58.6	4.0	41.8
CSP [31]	56.9	4.5	45.8
ALFNet+City [45]	56.8	4.5	43.4
SDS-RCNN [24]	56.8	6.4	38.7
MS-CNN [43]	55.8	9.5	48.6
BGCNet [74]	N.A.	4.1	42.0
TFAN+TDEM+PRM [76]	N.A.	6.7	30.9
DAGN [70]	46.8	6.0	33.2
PedHunter [77]	39.5	N.A.	N.A.
CSP+City [31]	54.4	3.8	36.5
CFRLA-Net (ours)	52.1	3.1	36.4
CFRLA-Net+City (ours)	50.9	2.1	25.3

To achieve faster detection speed and reduce the memory consumption of the MHSA path, we use the original 640×1280 images for testing instead of the 1024×2048 images used by most comparison methods. The result shows that the speed of our CFRLA-Net is 0.14 s/img, which is faster than the other state-of-the-art pedestrian detectors. This shows that our method can achieve state-of-the-art performance with lower resolution input, which further confirms the effectiveness of our method.

2) Caltech

We list the MR^{-2} performance of CFRLA-Net and some state-of-the-art methods [14, 24, 28, 31, 32, 43, 45, 74, 76, 77, 80] on the three subsets in Table V. Like some other methods (*e.g.*, ALFNet [45], OR-CNN [33], RepLoss [32], and CSP [31]), we also validate our detector with the model pre-trained on CityPersons [14]. As we can see, our CFRLA-Net+City achieves the best results and shows significant improvement over various existing methods on the three subsets. On the Heavy Occlusion subset, CFRLA-Net+City greatly pushes the performance to 25.3% MR^{-2}, which is over 11.2% better than the closest competitor (36.5% of CSP+City). And on the Partial Occlusion subset, CFRLA-Net+City also exhibits

the best performance, which is over 1.7% better than the closest competitor (3.8% of CSP+City). Even our CFRLA-Net is trained only with the training set of Caltech, the proposed detector still outperforms all the competing methods on the Partial Occlusion subset, and also presents competitive performance compared to the detectors pre-trained on CityPersons. These results firmly demonstrate the effectiveness of our detector for capturing the semantic and spatial context of occluded pedestrians.

3) CrowdHuman

Different from the CityPersons and Caltech datasets, the most crowded dataset CrowdHuman does not provide any occlusion subsets. Thus, we cannot get the performance of our method for different occlusion degrees of this dataset. To verify that our method can effectively deal with the CrowdHuman dataset, we compare our CFRLA-Net with some latest state-of-the-art methods with specific occlusion-handling strategies [32, 34, 35, 37, 38, 72, 78, 79, 82, 83] on the CrowdHuman validation set in Table VI. We can see that CFRLA-Net delivers the best performance in terms of Recall (96.7%) and AP (89.9%). Our detector using only the simple greedy-NMS algorithm still achieves comparable MR^{-2} (44.8%) compared to PBM+R^{2}NMS [34] and NOH-NMS [35] using some additional NMS strategies. Although the performance of our method is slightly lower than the two-stage anchor-based method Faster R-CNN+AEVB [82] using the AEVB occlusion-handling strategy, it is mainly due to the advantages of the anchor-based architecture over the anchor-free architecture. Compared with the anchor-free method FCOS+AEVB [82] using the AEVB occlusion-handling strategy, our method achieves a performance gain of 2.9% MR^{-2}. The relatively promising detection results in CrowdHuman also confirm the efficacy of our detector to extract the deep semantic context and precise spatial context, which can make the detector more robust to handle occlusions even without any specific occlusion-handling strategies.

Despite our method achieving good results in CrowdHuman, we have pointed out that it still has limitations when the pedestrian scenes are extremely crowded as analyzed in Sec. V. In the future, we will adopt a specific occlusion-handling strategy to improve the performance of our detector for heavily occluded pedestrians.

TABLE VI: Performance of CFRLA-Net in comparison with some state-of-the-art methods on CrowdHuman.

Method	MR^{-2}	Recall	AP
FPN [72]	52.4	90.6	83.1
FPN+Soft-NMS [38]	52.0	91.7	83.9
FPN+AdaptiveNMS [37]	49.7	91.3	84.7
RFB-Net [78]	65.2	94.1	78.3
RFB-Net+Soft-NMS [38]	66.3	95.4	78.1
RFB-Net+AdaptiveNMS [37]	63.0	94.8	79.7
GossipNet [79]	49.4	N.A.	80.4
RelationNet [83]	48.2	N.A.	81.6
Repulsion Loss [32]	45.7	88.4	85.6
PBM+R^{2}NMS[34]	43.4	93.3	89.3
NOH-NMS [35]	43.9	92.9	89.0
FCOS+AEVB [82]	47.7	N.A.	N.A.
Faster R-CNN+AEVB [82]	40.7	N.A.	N.A.
CFRLA-Net (ours)	44.8	96.7	89.9

V. Limitation and Discussion

During the visual analysis of the experimental results, we found that our detector missed some difficult pedestrian instances for some very crowded pedestrian scenes. As shown in Fig. 6, blue arrows represent pedestrian instances missed by our detector, which means that it seems difficult to obtain the discriminable features of these difficult pedestrian instances by extracting their complete context information when the occlusion is too severe. In the future, we expect to use the visible parts of human bodies provided in many modern pedestrian datasets [13–15] to extract the context information of these difficult pedestrian instances to enhance the discriminability of their features.

VI. Conclusion

In this paper, we propose a simple but effective CFRLA-Net for pedestrian detection in crowd scenes. Based on the anchor-free detection framework, we introduce a CFRL-Module, the core of which is a CEPCA-Block with two parallel paths. The CEPCA-Block integrates convolution and MHSA with low computational cost, which can obtain the deep semantic context by the convolution path and precise context by the MHSA path. Furthermore, we propose a novel local window MHSA with a specific aspect ratio window to achieve an efficient balance between performance and memory consumption in high-resolution pedestrian detection. By parsing the context features generated by CFRL-Module via the anchor-free detection head, our CFRLA-Net can catch a high-level understanding of the heavily occluded and small-scale pedestrian instances based on HRNet, which can effectively solve the limitation of the insufficient feature extraction ability of HRNet for the hard samples. Extensive experiments on three challenging pedestrian detection benchmarks (*i.e.*, CityPersons, Caltech, and CrowdHuman) demonstrate that our detector achieves state-of-the-art results, and strongly validates the superiority of the proposed CFRLA-Net.

References

[1] T. Liu, K. Lam, R. Zhao, and G. Qiu, "Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 315-329, 2021.

[2] L. Chen, S. Lin, X. Lu, D. Cao, and F. Y. Wang, "Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234-3246, 2021.

[3] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, and C. Fox, "Pedestrian models for autonomous driving part I: Low-level models, from Sensingto tracking," *IEEE Trans. Intell. Transp. Syst.*, pp. 1-21, 2020.

[4] F. Bu, T. Le, X. Du, R. Vasudevan, and M. Johnson-Roberson, "Pedestrian Planar LiDAR Pose (PPLP) Network for Oriented Pedestrian Detection Based on Planar LiDAR and Monocular Images," *IEEE Robot. Auto. Letters*, vol. 5, no. 2, pp.1626-1633, 2020.

[5] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 1243-1257, 2015.

[6] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1140-1151, 2008.

[7] W. Si, H. S. Wong, and S. Wang, "Variant semiboost for improving human detection in application scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 1595-1608, 2018.

[8] X. Wang, C. Liang, C. Chen, Z. Wang, and J. Chen, "S^3D: Scalable pedestrian detection via score scale surface discrimination," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3332-3344, 2020.

[9] K. Chen and Z. Zhang, "Pedestrian counting with back-propagated information and target drift remedy," *IEEE Trans. Cybern.*, vol. 47, no.4, pp. 639-647, 2016.

[10] M. Bilal, A. Khan, M. U. K. Khan, and C. M. Kyung, "A low complexity pedestrian detection framework for smart video surveillance systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2260 C 2273, 2017.

[11] C. Lin, J. Lu, and J. Zhou, "Multi-Grained deep feature learning for robust pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3608-3621, 2019.

[12] Y. Jiao, H. Yao, and C. Xu, "PEN: Pose-Embedding network for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1150-1162, 2021.

[13] P. Dollr, 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, 2012.

[14] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3213-3221, 2017.

[15] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.

[16] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, An empirical study of context in object detection, in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1271-1278, 2009.

[17] R. Mottaghi, X. Chen, X. Liu, and N.G. Cho, et al, "The role of context for object detection and semantic segmentation in the wild," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 891-898, 2014.

[18] D. Parikh, C.L. Zitnick, and T. Chen, "Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1978-1991, 2012.

[19] X. Chang, P. Huang, Y. Shen, X. Liang, Y. Yang, and A.G. Hauptmann, "RCAA: Relational context-aware agents for person search," in *Proc. Eur. Conf. Comput. Vis.*, pp. 84-100, 2018.

[20] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 966-974, 2018.

[21] 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 Int. Conf. Comput. Vis.*, pp. 4967-4975, 2019.

[22] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6995-7003, 2018.

[23] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3486-3495, 2017.

[24] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis.*, pp. 135-151, 2018.

[25] M. Xu, Y. Bai, S. S. Qu, and B. Ghanem, "Semantic part rcnn for real-world pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 45-54, 2019.

[26] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1904-1912, 2015.

[27] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. Davis, "Fused deep neural networks for efficient pedestrian detection," *arXiv preprint arXiv:1805.08688*, 2018.

[28] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis.*, pp. 443-457, 2016.

[29] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985-996, Oct. 2017.

[30] Q. Li, Y. Su, Y. Gao, F. Xie, and J. Li, "OAF-Net: An occlusion-aware anchor-free network for pedestrian detection in a crowd," *IEEE Trans. Intell. Transp. Syst.*, pp. 1-10, 2022.

[31] 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 Recog.*, pp. 5187-5196, 2019.

[32] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7774-7783, 2018.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis.*, pp. 637-653, 2018.

[34] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10750-10759, 2020.

[35] P. Zhou, C. Zhou, P. Peng, J. Du, X. Sun, X. Guo, and F. Huang, "NOH-NMS: Improving pedestrian detection by nearby objects hallucination," in *Proc. 28th ACM Int. Conf. Multimedia*, pp. 1967-1975, 2020.

[36] Q. Li, Y. Bi, R. cai, and J. Li, "Occluded Pedestrian Detection through Bi-Center Prediction in Anchor-Free Network," *Neurocomputing*, pp. 199-207, 2022.

[37] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6459-6468, 2019.

[38] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 5562-5570, 2017.

[39] J. Ren, X. Chen, J. Liu, W. Sun, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5420-5428, 2017.

[40] H. Zhang, K. Wang, Y. Tian, C. Gou, and F. Y. Wang, "Mfr-cnn: Incorporating multi-scale features and global information for traffic object detection," *IEEE Trans. Vehicu. Technol.*, vol.67, no.9, pp. 8019-8030, 2018.

[41] C. Fei, B. Liu, Z. Chen, and N. Yu, "Learning pixel-level and instance-level context-aware features for pedestrian detection in crowds," *IEEE Access*, vol.7, pp. 94944-94953, 2019.

[42] 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.*, pp. 21-37, 2016.

[43] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, pp. 354-370, 2016.

[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, Jun. 2012.

[45] 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.*, pp. 618-634, 2018.

[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 779-788, 2016.

[47] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 850-859, 2019.

[48] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis.*, pp. 734-750, 2018.

[49] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9657-9666, 2019.

[50] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6569-6578, 2019.

[51] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[52] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," *arXiv preprint arXiv:1904.03797*, 2019.

[53] Q. Li, H. Qiang, and J. Li, "Conditional random fields as message passing mechanism in anchor-free network for multi-scale pedestrian detection," *Info. Sci.*, vol. 550, pp. 1-12, Oct. 2021.

[54] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, and N. Golmant, et al, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9127-9135, 2018.

[55] W. Chen, D. Xie, Y. Zhang, S. Pu, "All you need is a few shifts: Designing efficient convolutional neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7241-7250, 2019.

[56] H. You, X. Chen, Y. Zhang, C. Li, and S. Li, et al, "ShiftAddNet: A hardware-inspired deep network," *Proc. Adv. Neural Inf. Process. Syst.*, vol.33, pp. 2771-2783, 2020

[57] Q. Chen, Q. Wu, J. Wang, and Q. Hu, et al, "MixFormer: Mixing features

across windows and dimensions," *arXiv preprint arXiv:2204.02557*, 2022.
[58] X. Pan, C. Ge, R. Lu, and S. Song, et al, "On the Integration of Self-Attention and Convolution," *arXiv preprint arXiv:2111.14556*, 2021.
[59] W. Wang, E. Xie, X. Li, D. P. Fan, and K. Song, et al, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.
[60] Z. Liu, Y. Lin, Y. Cao, and H. Hu, et al, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 10012-10022, 2021.
[61] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.
[62] A. Srinivas, T.Y. Lin, N. Parmar, J. Shlens, and P. Abbeel, et al, "Bottleneck Transformers for Visual Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16519-16529, 2021.
[63] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 248-255, 2009.
[64] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5693-5703, 2019.
[65] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9557-9566, 2019.
[66] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318-327, Feb. 2020.
[67] J. Zhang, L. Lin, J. Zhu, Y. Li, and C. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Trans. Multimedia*, pp. 1-13, Sep. 2020.
[68] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1259-1267, 2016.
[69] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[70] H. Xie, W. Zheng, and H. Shin, "Occluded pedestrian detection techniques by deformable attention-guided network," *Appl. Sci.*, vol. 11, no. 13, pp. 1-19, Jun. 2021.
[71] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944-954, May, 2017.
[72] T. Y. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2117-2125, 2017.
[73] T. Song, L. 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.*, pp. 536-551, 2018.
[74] J. Li, S. Liao, H. Jiang, and L. Shao, "Box guided convolution,pedestrian detection,receptive fields,scale variation," in *Proc. 28th ACM Int. Conf. Multimedia*, pp. 1615-1624, 2020.
[75] X. Song, K. Zhao, W. S. Chu, H. Zhang, and J. Guo, "Progressive refinement network for occluded pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, pp. 32-48, 2020.
[76] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, and J, Yuan, "Temporal-context enhanced detection of heavily occluded pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13427-13436, 2020.
[77] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "PedHunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. IEEE Conf. Comput. Artif. Intell.*, pp. 10639-10646, 2020.
[78] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, pp. 404-419, 2018.
[79] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4507-4515, 2017.
[80] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3127-3136, 2017.
[81] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7231-7240, 2019.
[82] Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, "Variational pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 11617-11626, 2021.
[83] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3588-3597, 2018.

Jun Li (M'18) is currently the director of Fujian robotics intelligent system engineering technology research center. He earned his Ph.D. from University of Munich, Germany. He was a postdoctoral research associate at University of Marburg, Germany. In 2015, he joined the faculty at Haixi Institute, Chinese Academy of Sciences. His research focuses on the image processing and recognition, robot adaptive control, and human-computer interaction. In the field of images and robots, Prof. Jun Li has published two personal English and German monograph, published more than 40 innovative academic papers in IEEE and other SCI journals, applied more than 30 invention patents, and obtained 3 National invention patents. In 2016, he was selected as member of the Fujian Provincial Double-Hundred-Talent Program. Since 2015, prof. Jun Li as the leader of the Laboratory of Robotics and Intelligent Systems hosted and participated different scientific research programs, including the national 13th Five-Year Key Research Program and Fujian major science and technology projects.

YuQuan Bi received the B.S. degree in engineering from Beijing Forestry University, China. He is currently pursuing the M.Eng degree with the department of advanced manufacturing, Fuzhou University, China. His research interests include computer vision, object detection and deep learning.

Sumei Wang received the B.Eng. degree from Hebei University, China in 2013, and the Ph.D. degree from Zhejiang University, China in 2018. She is currently a Postdoctoral Fellow at The Hong Kong Polytechnic University, Hong Kong. Her research interests include structural dynamics, rail-bridge interaction, machine vision, and structural health monitoring.

Qiming Li is currently an associate research fellow in Haixi Institutes, Chinese Academy of Sciences. He received the PhD degree in the School of Information Science and Technology from Xiamen University, in 2016. His research interests include computer vision, object detection and tracking, and machine learning. He has presided over several projects at various levels such as National Natural Science Foundation of China, China Post-doctoral Science Foundation, and many other provincial-level projects. He has authored over 20 SCI indexed journal papers and EI indexed refereed conference papers including *IEEE Transactions on Cybernetics*, *IEEE Transactions on Industrial Electronics*, *IEEE Transactions on Intelligent Transportation Systems*, *Signal Processing*, etc.