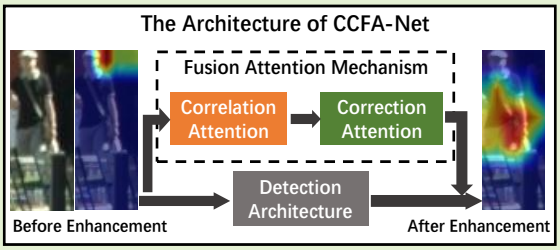
Correlation-and-Correction Fusion Attention Network for Occluded Pedestrian Detection

Fengmin Zou, Xu Li, *Member, IEEE*, Qimin Xu, Zhengliang Sun, and Jianxiao Zhu

***Abstract*—As a significant issue in computer vision, pedestrian detection has achieved certain achievements with the support of deep learning. However, pedestrian detection in congested scenes still encounters the challenging problem of feature loss and obfuscation. To address the issue, we propose a pedestrian detection network based on a correlation-and-correction fusion attention mechanism. First, a multi-mask correction attention module is proposed, which generates visible part masks of pedestrians, enhancing the visible region's features and correcting the false one. Besides, the module preserves the features of multi-class pedestrians by generating multiple masks. Then, we fuse a correlation channel attention module to enhance the correlation of various pedestrians' body features. Next, we studied three fusion methods of correlation and correction attention mechanisms and found that the serial connection of "correlation first and correction behind" works best. Finally, we extend our method to multi-class pedestrian detection in congested scenes. Experimental results on the CityPersons, Caltech and CrowdHuman datasets demonstrate the effectiveness of our method. On the CityPersons dataset where more than 70% of pedestrians are occluded, our method outperforms the baseline method by 1.12% on the heavy occlusion subset and surpasses many outstanding methods.**



***Index Terms*— Attention mechanism, crowded scenes, feature enhancement, fusion, pedestrian detection**

I. Introduction

With the growth of vehicle usage worldwide, traffic safety hazards are constantly emerging. To solve this problem, many studies have been carried out[1]-[3], among which, the Intelligent Vehicle Infrastructure Cooperative System provides a new idea. It uses communication technology to connect vehicles and the surrounding environment, and realize sufficient information interaction between vehicles, roads and traffic participants, thereby effectively improving transportation efficiency, alleviating traffic congestion, and reducing the frequency of traffic accidents. As an essential component of the system, the environmental perception system undertakes the crucial task of reliable and accurate perception of various traffic participation elements. Among them, pedestrians, as one of the core traffic participation factors, are seriously affected by the uncertainty and randomness of the traffic condition, and are prone to congestion in complex traffic environments, thus bringing huge technical challenges to accurate environmental perception.

Pedestrian detection based on RGB images from the vision sensor has become a hot issue, which is widely used in autonomous driving, traffic monitoring, and many other fields. Research based on Convolutional Neural Networks (CNN) has made significant achievements[4]-[8]. However, due to the complexity of the scene and pedestrian poses, high-precision detection is still a struggle among which the crowded scene is extremely challenging.

In a crowd, pedestrians' body parts are covered, making it tough for CNN to accurately extract the characteristics of pedestrians. There are some studies proposing solutions. Some divide the human body into parts and learn the features separately[9],[10], but how to combine these features can be quite an intricacy[11]. Some introduce segmentation information for the human head or visible parts to enhance features[12]-[14], however, numerous annotations result in extra calculations. And others focus on improving the non-maximum suppression (NMS) strategy[15]-[17] which is limited to intra-class occlusion.

Different from the above methods, we propose the correlation-and-correction fusion attention network (CCFA-Net), which adds serially fused attention modules to the basic detection architecture to enhance effective features in a crowd. First, CCFA-Net employs a correlation channel attention module to explicitly model interdependences of channels in

Manuscript received xxx; revised xxx; accepted xxx. Date of publication xxx; date of current version xxx. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3904403, in part by the National Natural Science Foundation of China under Grants 61973079, and in part by the Primary Research & Development Plan of Jiangsu Province, grant number BE2022053-5, BE2019106. The associate editor coordinating the review of this article and approving it for publication was xxx (Corresponding author: Xu Li).

Fengmin Zou, Xu Li, Qimin Xu and Jianxiao Zhu are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: m15840593711@163.com; lixu.mail@163.com; Jimmy.xqm@gmail.com; jianxiao_zhu@seu.edu.cn).

Zhengliang Sun is with the Traffic Management Research Institute, Ministry of Public Security, Wuxi 214151, China (e-mail: szl8205@sina.com).

Digital Object Identifier xxx.

diverse occlusion patterns since different channels activate responses for different body parts[18]. This module emphasizes the feature of visible body parts and suppresses the occluded ones by establishing and enhancing the channel-wise correlation in multiple occlusion patterns. Then, CCFA-Net fuses a multi-mask correction attention module, which introduces external supervision of visible regions from the spatial dimension to perform secondary feature recalibration. The module re-emphasizes the feature of visible parts while correcting the false enhancement caused by correlation attention in complex intra-class occlusion mode by capturing the spatial dependencies of each position in feature maps. Besides, it generates multi-class pedestrian masks, enabling the network to perform multi-class pedestrian detection in crowded, which is significant for application in complicated traffic scenarios. The organic fusion of these two attention mechanisms enables the network to effectively solve the multiple complex occlusion patterns including intra-class and inter-class.

Contribution. our main contributions can be summarized as follows:

1) We propose a novel pedestrian detection network based on a correlation-and-correction fusion attention mechanism resolving the multi-pattern occlusion problem by feature recalibration.
2) The correlation-and-correction fusion attention module (CCFA) is proposed and the fusion methods are studied. The fusion module consists of the correlation attention and the correction attention, emphasizing the effective feature of visible parts while suppressing the occluded ones. And we find that the serial fusion method of " correlation first and correction behind" achieves the best improvement effect.
3) A multi-mask correction attention module is proposed to explicitly enhance the visible region features and correct the false enhancement. Remarkably, we extend CCFA-Net to multi-class occluded pedestrian detection by generating the multi-class masks in this module that retain the characteristics of multi-class pedestrians.
4) The proposed network achieves excellent results on CityPersons[4] , Caltech[19] and CrowdHuman[20] datasets. Especially on the heavy occlusion subset of CityPersons[4], CCFA-Net outperforms most outstanding methods. Furthermore, the experiments of CCFA-Net on our collected data demonstrate its extendibility to the multi-class pedestrian detection problem.

II. Related Work

A. Pedestrian Detection

In recent years, CNN-based methods have made significant progress in pedestrian detection[21]-[24], which can be divided into one-stage and two-stage methods. The two-stage method generates region proposals before classification and regression, whereas the one-stage method obtains the results with only one step to improve the detection speed regardless of the decline in accuracy. Therefore, to ensure accuracy in complex scenes, we adopt the two-stage method.

Early methods of using CNN in pedestrian detection are mostly based on RCNN[25] method. With the proposal of the Region Proposal Network (RPN), Faster RCNN[26] has become the primary two-stage network applied by most pedestrian detection methods[27]-[29]. And some studies improve it for pedestrian targets. For example, [30]proposes five modifications to the general Faster RCNN[26] to better handle small persons; Reference [31]combined SRGAN[32] and improved Faster RCNN[26] to detect pedestrians in low-quality images; The multi-scale framework proposed by MS CNN[33] is appropriate for detecting pedestrians of different scales. Since the two-stage methods present excellent performance, we still adopt it as our detection architecture and apply Faster RCNN[26] as the basic structure of the network.

B. Occlusion Handling

Although CNN-based methods perform well in the general scene, detecting in a crowd is still quite challenging. Recently, some articles have researched this issue, which can be summarized into three approaches. The first is to improve the detector based on pedestrian body parts[9],[10],[34]. This method divides the whole body into multiple parts to train the detectors of them separately for different occlusion modes. The difficulty is how to effectively combine these various detectors. And the numerous detectors cause computational resource consumption and slowing down of detection speed. The second approach improves the non-maximum suppression (NMS) strategy when conducting post-processing[15],[17],[35]. It improves the detection errors caused by the overlapping of region proposals in a crowd, but it's unremarkable for inter-class occlusion. The third one is based on the attention mechanism[13],[14],[18], which insets an attention module into the general detection network to help it focus on the local parts of the features. Research that utilizes attention mechanisms either uses self-attention or introduces head or visible parts as guidance. Different from them, we efficiently fuse correlation channel attention and multi-mask correction attention to achieve excellent results.

C. Attention Mechanisms

Attention mechanism has been commonly applied in various fields of computer vision[36]-[38], whose effectiveness stems from the imitation of human vision that automatically captures local key features when observing the whole object. There are some studies using attention mechanisms to help detect pedestrians. Reference [13] utilizes the location information of the visible area as an additional guide for the network to pay attention to the visible part of the human body. Reference [14] employs the human head information to adjust the focus area. However, the above methods ignore the significant relationship between channel characteristics and pedestrian body parts [18]. Reference [39] proposes the self-attention model SENet to recalibrate channel-wise feature responses, but it ignores the relevance of features in the spatial dimension.

Different from the above research, we simultaneously fuse the correlation channel attention and the spatial multi-mask correction attention in an efficient structure and study its effective fusion method, so that the network can effectively

solve multi-pattern occlusions through feature recalibration.

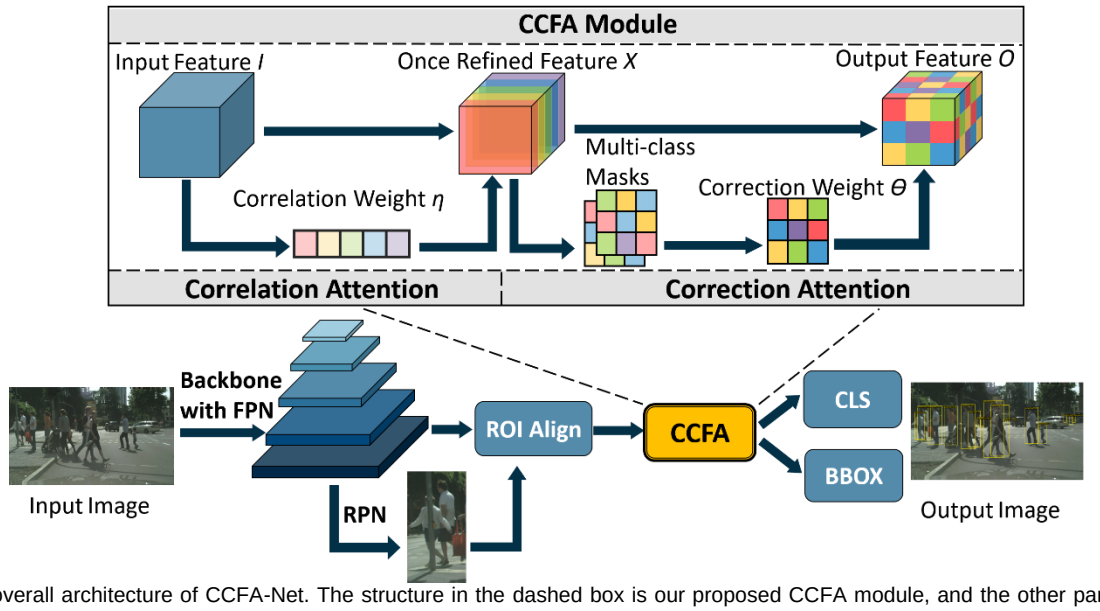


Fig. 1. The overall architecture of CCFA-Net. The structure in the dashed box is our proposed CCFA module, and the other parts are the basic detection architecture Faster RCNN, which together constitute CCFA-Net. The input image is extracted by the backbone, and then the proposal feature map is generated by RPN and ROI Align. Then it is classified and regressed after two re-weightings by CCFA to obtain the final output.

III. Proposed Method

We propose a pedestrian detection network based on a correlation-and-correction fusion attention mechanism. As shown in Fig. 1, CCFA-Net consists of the general detection architecture Faster RCNN[26] and a fusion attention module which contains correlation and correction attention mechanisms.

As an effective basic detection architecture, Faster RCNN[26] can perform excellently in general scenes. However, it is problematic to extract features in a crowd since the physical characteristics of pedestrians are occluded by other objects, resulting in severe detection errors. To tackle this problem, we propose CCFA, which weights the features twice from two dimensions to enhance the effective body part features and the visible region features of pedestrians so that the intra-class and inter-class occlusion problem is resolved. Next, we will describe the overall detection architecture, attention modules, and the fusion methods of CCFA-Net in detail.

A. Detection Architecture

The detection methods in pedestrian detection are divided into one-stage and two-stage approaches. Compared with the one-stage method, the two-stage way achieves high detection accuracy. As the representative network in the two-stage methods, Faster RCNN[26] performs excellently. Therefore, we employ it as our basic detection architecture.

The detection pipeline is shown in Fig. 1. First, the raw image is input into the ResNet-50 classification network pre-trained on ImageNet for feature extraction. To enhance the multi-scale adaptability of the network, we add a Feature Pyramid Network (FPN)[40] after ResNet-50 to obtain the 5-layer feature maps to fuse multi-scale features. Then, RPN generates region proposals and maps them to the 5-layer feature maps to obtain proposal feature maps. Next, the ROI Align layer extracts the proposal features and resizes them to a fixed size. Finally, the proposals are classified and regressed through fully connected layers. The network is optimized through the loss function:

$$L_0 = L_{rpn\_cls} + L_{rpn\_reg} + L_{rcnn\_cls} + L_{rcnn\_reg} \quad (1)$$

Where L_{rpn_cls} and L_{rcnn_cls} are the cross-entropy classification loss functions, and L_{rpn_reg} and L_{rcnn_reg} are the L_1 regression loss functions.

Although Faster RCNN[26] performs well in general scenes, there is still a serious problem in the crowd. This issue occurs due to the effective features that can be used to identify pedestrians are occluded, which causes worse results. Therefore, we propose the fusion attention mechanism module as shown in Fig. 1, and embed it into the detection process to re-weight the feature, thus reducing the influence of the occluded area on the detection result.

Fig. 2. Different channels activate responses for different body parts. Highlighted regions represent strong activation of each representative channel.

B. Correlation Channel Attention Mechanism

The channel attention mechanism is used to enhance the crucial feature in an image since each channel in the feature map corresponds to a certain feature. By enhancing the correlation between the channels, the network can concentrate

on the features that have a decisive impact on the detection results. As far as pedestrian detection, the main features that distinguish pedestrians from other objects are the characteristics of each body part. Reference [18]proves that different channels activate responses for different body parts. As shown in Fig. 2, for two representative channels, one shows strong activation to the person's head and the other reacts to the person's legs. Therefore, we employ a channel attention mechanism to reduce misclassification caused by obscured body parts in a heavily crowded scene.

SENet[39] is a representative network applying the channel attention mechanism in CNN, which proposes the "Squeeze-and-Excitation" (SE) block to reweight the channel features. The SE block squeezes global spatial information of the input feature $I \in \mathbb{R}^{H\times W\times C}$ through a global average pooling layer to generate a statistic $z \in \mathbb{R}^{C}$:

$$z_c(I) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} I_c(i,j) \tag{2}$$

then, the weight vector $\eta \in \mathbb{R}^{C}$ is obtained through two fully connected layers. And the output feature $X \in \mathbb{R}^{H\times W\times C}$ is reweighted by the η:

$$g(I) = \sigma(w_2\delta(w_1 z(I))) \tag{3}$$

$$X = \eta \odot I = g(I) \odot I \tag{4}$$

where δ refers to ReLU function; σ refers to sigmoid function; $g(\bullet)$ represents the function that generates η, i.e. $g(I)$ equals to η; $\odot$ represents the channel-wise multiplication for each element between the three-dimensional feature map $I \in \mathbb{R}^{H\times W\times C}$ and the one-dimensional weight vector $\eta \in \mathbb{R}^{C}$. And the parameter $W_1 \in \mathbb{R}^{C/r\times C}$, $W_2 \in \mathbb{R}^{C\times C/r}$,which is the same as[37], and we set the reduction ratio r as 16.

The SE block models interdependencies between channels to readjust channel-wise features which activate responses to different body parts. Besides, the structure is simple and it does not change the dimension of input and output channels, which is lightweight and easy to be embedded in other structures. Therefore, we embed the SE block as a channel attention mechanism into the basic detection architecture, which adjusts the weight of each channel related to pedestrian body parts without consuming a lot of computing resources. Subsequent experiments demonstrate that the channel attention composed of SE block effectively reduces the missed detection in a crowded scene.

The essential role of this block is to fully establish and strengthen the correlation of each body part to adapt the detector to complex occlusion patterns by explicitly modeling channel-wise interdependencies, so we call this attention mechanism correlation channel attention.

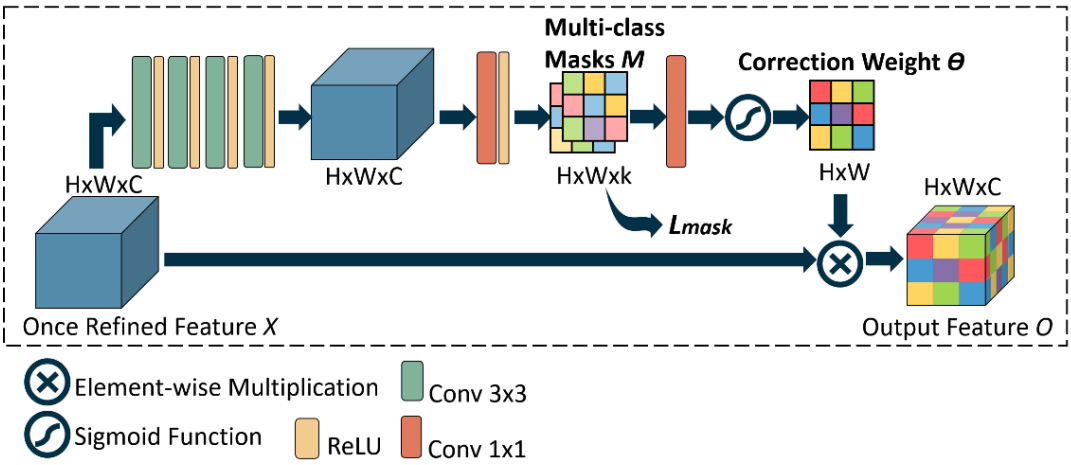


Fig. 3. The architecture of multi-mask correction attention module. The feature map weighted by correlation attention once is input into this module, which generates multi-class masks and correction weight, and the correction weight re-weights the input twice to obtain the output.

C. Multi-mask Correction Attention Mechanism

Different from the correlation channel attention module that generates the weights by shrinking the global feature map, the correction attention module obtains the weights by additionally introducing the segmentation information for the visible region, which captures the local spatial dependencies of pedestrians to enhance features of local key regions.

Mask RCNN[41] is a representative network for instance segmentation tasks, which adds parallel mask branches based on the structure of Faster RCNN[26]. Its research shows that adding the mask branch to Faster RCNN[26] can improve the detection result. Inspired by it, we propose a multi-mask correction attention module. As shown in Fig. 4, the input of the module is the feature map $X \in \mathbb{R}^{H\times W\times C}$ that has been reweighted once by the correlation attention module, where H, W, and C are the height, width, and channel dimension of X, respectively. Then four 3×3 convolution layers perform high-dimensional feature extraction on X, and the convolved feature map is passed through the 1×1 convolution kernel to obtain the multi-class masks $M \in \mathbb{R}^{H\times W\times k}$, in which k is the number of classification categories.

$$M(X) = f_1^{1\times 1}((f_1^{3\times 3}(X))_4) \tag{5}$$

where $f_1^{1\times 1}$ and $f_1^{3\times 3}$ represent the convolution operations of 1×1 and 3×3 with the ReLU activation function, respectively. Next, the mask is again subjected to a 1×1 convolution layer and the sigmoid function to obtain the

correction weight $\theta \in \mathbb{R}^{H\times W}$.

$$h(X) = \sigma(f^{1\times1}(M(X))) \quad (6)$$

$$\theta = h(X) \quad (7)$$

where $f^{1\times1}$ represents the convolution operations of 1×1, and $\sigma(\bullet)$ represents the sigmoid function. After that, θ is multiplied by the value of each position of each channel on the input feature map, that is, the feature map $O \in \mathbb{R}^{H\times W\times C}$ after the second weighting of the multi-mask correction attention mechanism is obtained:

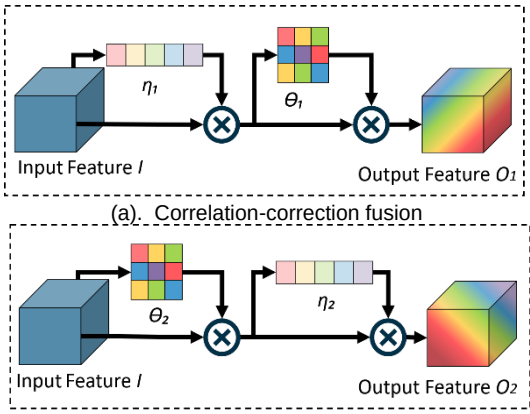
$$O = \theta * X = h(X) * X \quad (8)$$

where $*$ represents element-wise multiplication for each channel between the three-dimensional feature map $X \in \mathbb{R}^{H\times W\times C}$ and the two-dimensional weight matrix $\theta \in \mathbb{R}^{H\times W}$. $H\times W\times C$ is set to $7\times7\times256$. In this process, the weights are optimized by an averaged binary cross-entropy loss function L_{mask}, which is defined the same as [41].

Unlike most pedestrian detection methods where spatial attention only focuses on single-class pedestrian detection, we also pay attention to the multi-class task since there are diverse person states in traffic scenes such as walking, cycling, sitting, and so on. Therefore, we propose two 1×1 convolution layers to adapt to multi-class detection. The first one generates multiple masks for multiple classes and thus retains multi-class features, and the second one performs channel compression to generate the correction weight. In this way, the network performs remarkably in occluded scenes even when executing multi-class pedestrian detection.

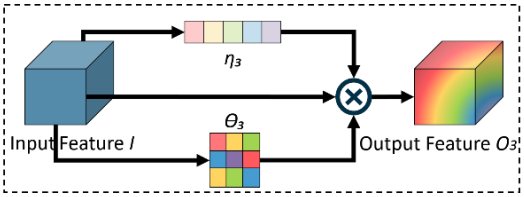
This module needs to introduce additional segmentation information during the training process. In order to reduce the complexity of labeling data and enhance the applicability of our method on various datasets, we use the coarse-level segmentation information of the visible part. Particularly, we use the visible region bounding-box rather than the semantic segmentation annotations to generate the mask.

Note that we call this module correction attention since it can correct the "false enhancement" caused by correlation attention as shown in Fig. 6. And we will illustrate it in detail later.



(a). Correlation-correction fusion

(b). Correction-correlation fusion

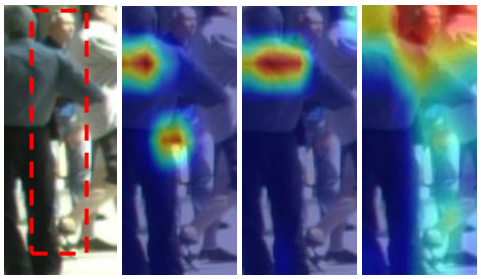


(c). Fusion in parallel

Fig. 4. The fusion methods of correlation and correction attention modules. (a) is the serial fusion method of correlation attention first and correction attention behind; (b) is the serial fusion method of correction attention first and correlation attention behind; (c) shows the parallel fusion method of correlation and correction attention.

(a) (b) (c) (d)

Fig. 5. Progressive enhancement of features by CCFA. (a) is the original image where the detection target is in the red box. (b) is the activation map of the ROI feature. (c) is the activation map of the feature re-weighted only by correlation attention, and (d) shows the activation map re-weighted again by correction attention. The degree of re-weighting for target features by the two attention mechanisms increases gradually.



(a) (b) (c) (d)

Fig. 6. "False enhancement" caused by correlation attention and the corrective effect of correction attention. (a)-(d) are the same as in Fig. 5.

D. Fusion of Attention Modules

The correlation attention module employs global average pooling to shrink the features through spatial dimension, essentially ignoring the interdependencies of each local area in space, while the correction attention module generates the weight through 1×1 convolution, which compresses the channel features, ignoring the interaction between local features and the global receptive fields. Both of them re-weight the features of one dimension and lose the features of another, so it is difficult for a single attention mechanism to adapt to various scenes with different degrees of occlusion. Therefore, we organically fuse the correlation and correction attention mechanisms to form an effective fusion attention module.

We studied three fusion methods of the two attention

mechanisms. The structure is shown in Fig. 3. The three structures represent the 'correlation-correction' serial fusion, the 'correction-correlation' serial fusion, and the connection in parallel. The three structures all employ the correlation attention weight vector η and the correction attention weight matrix θ to adjust the attention to the input features I :

$$O_n = I \odot \eta_n * \theta_n \tag{9}$$

where n=1,2,3. And different fusion methods will generate different η and θ :

$$\begin{cases} \eta_1 = g(I) \\ \theta_1 = h(\eta_1 \odot I) \end{cases} \tag{10}$$

$$\begin{cases} \eta_2 = g(\theta_2 * I) \\ \theta_2 = h(I) \end{cases} \tag{11}$$

$$\begin{cases} \eta_3 = g(I) \\ \theta_3 = h(I) \end{cases} \tag{12}$$

where the definitions of $g(\bullet)$ and $h(\bullet)$ are the same as (3) and (6).

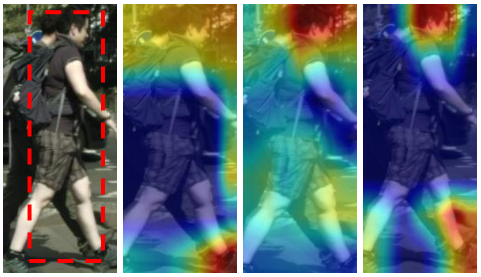
It can be seen from the above formulas that different attention fusion methods will generate different weights that affect the output features. We adopt the 'correlation-correction' serial fusion method. For one thing, fusion in this way enables progressive feature enhancement. As shown in Fig. 5, the attention to key features is gradually transferred from the occluded leg region to the non-occluded target head region which is beneficial to the detection results. For another thing, we found that correlation attention brings 'false enhancement' in the case of extremely severe intra-class occlusion. As shown in Fig. 6, the inter-class occlusion between pedestrians is often caused by the interleaving of body parts, which leads to the correlation attention module enhancing the false body part that does not belong to the target pedestrian, thus resulting in the 'false enhancement', whereas the external supervision of the correction attention module can reduce this phenomenon, achieving the effect of correction. Therefore, adopting the 'correlation-correction' serial fusion method enables the module to comprehensively achieve progressive feature enhancement, covering multiple occlusion patterns from inter-class to intra-class. Fig. 7 visualizes the improvement effect of CCFA by Grad-CAM[42].

Improper fusion methods may cause the attention to fail to fully exert its feature enhancement effect, and even play a reverse modulation effect to make the result worse. For example, the 'correction-correlation' fusion method cannot take advantage of the strong feature correction effect of the correction attention module, and the subsequent correlation attention is prone to the 'false enhancement' when pedestrians occlude each other severely, which makes the network unable to adapt to the severe intra-class occlusion as shown in Fig. 6. Besides, the parallel fusion method separates the two attentions from each other, obtains weights separately and then modulates the original features together. In this way, the two attentions cannot capitalize mutual influence, and their shortcomings of losing a certain dimension feature are amplified, which leads to the reverse feature modulation effect of the fused attention, making the effect worse. As shown in Fig. 8, the parallel fusion method in Fig.8 (b) leads the network pay too much attention to a large area around the pedestrian target, bringing focus on non-targets when occluded severely. In contrast, the serial fusion method in Fig.8 (c) and (d) can exactly focus on the target, and the 'correlation-correction' serial fusion method in (d) focuses more clearly on the key parts of the target.

Furtherly, the following experimental results show that the network structure that integrates the two attention mechanisms in 'correlation-correction' serial fusion method outperforms others.

(a) (b) (c) (d) (e) (f)

Fig. 7. Visualization of attention to features before and after employing the CCFA module. (a) and (d) are the original image where the detection target is in the red box. (b) and (e) show the activation map of each ROI feature, and (c) and (f) are the activation maps of the features modulated by the CCFA module.



(a) (b) (c) (d)

Fig. 8. Visualization of attention weights learned in three fusion methods. (a) is the original image where the detection target is in the red box. (b) shows the parallel fusion method, (c) shows the 'correction-correlation' serial fusion method and (d) shows the 'correlation-correction' serial fusion method which is adopted in CCFA module.

IV. Experiments

To evaluate our proposed method, we conduct experiments on CityPersons, Caltech and CrowdHuman datasets, and the experimental results demonstrate the effectiveness of our proposed CCFA-Net. In the next subsections, we will first introduce the datasets, evaluation metrics, and specific experimental details, and then introduce the results of the ablation experiments and advanced comparison experiments. Finally, we will extend CCFA-Net to a multi-class pedestrian detection task and show the detection effect.

A. Datasets

CityPersons. The CityPersons dataset[30] is a subset of the CityScapes dataset[43] generated by secondary annotation of pedestrians. It contains 5000 urban road scene images of 27 cities in Germany, among which the training, validation, and test set are composed of 2975, 500, and 1575 images, respectively. There are numerous diverse crowded scenes in

the CityPersons dataset, which not only provides high-quality full-body annotations for pedestrians but also marks the visible body areas for pedestrians. Therefore, most of our experiments will be carried out based on it. Since the CityPersons dataset does not disclose the annotation information of the test set, we test our model on the validation set.

Caltech. The Caltech dataset[19] is one of the most commonly used datasets in pedestrian detection , which consists of 10 hours of urban traffic environment in-vehicle video. The video with a size of 640×480 and a frequency of 30hz contains about 250,000 frames, where a total of 2,300 pedestrians are labeled, and 350,000 labeled boxes are generated, which contain the full body and the visible parts annotations of the pedestrians. All the data in the Caltech dataset are contained in 11 videos, of which the 0-5 videos are used as the training set, and the 6-10 videos are used as the test set. We sample the training set and the test set at a sampling frequency of 10 Hz and 1 Hz, respectively, and generated a total of 42,782 images for training and 4024 images for testing. Reference [5] relabeled the Caltech dataset to improve the labeling quality of the original dataset. Therefore, we use Caltech's new annotations in our experiments.

CrowdHuman. The CrowdHuman dataset is a large and diverse dataset that focuses on intra-class occlusion of people. The training and validation subsets contain a total of 470K human instances with a variety of occlusion patterns. The training, validation and test subsets each contain 15000, 4370 and 5000 images. Each human instance is annotated with a head bounding-box, human visible-region bounding-box and human full-body bounding-box. Note that we only need the visible-region box and the full-body box. Since the annotations of testing subset are not publicly available, we train the models on the training subset and evaluate on the validation subset.

B. Evaluation Metrics

The standard average-log miss rate (MR) is a commonly used evaluation metric in pedestrian detection. We use MR computed in the false positive per image (FPPI) range of [10^{-2}, 10^{0}][19], denoted as MR^{-2}, to evaluate the results of all the experiments, where a lower number is better. Since we are more concerned about scenes with severe occlusion, we will test the model in three different occlusion modes with visibility in the range of [0, 0.90], which are denoted as "Partial", "Heavy" and "All" subsets. In addition, since various pedestrian detection methods have different classification standards for occlusion modes[5],[6], to facilitate subsequent comparisons of the state-of-the-art, we add a fourth occlusion mode "Heavy*". The visibility and height of the data in the four subsets are shown in Table I.

TABLE I
VISIBILITY AND HEIGHT OF FOUR OCCLUSION MODES OF SUBSETS.

Subset	Visibility	Height
Partial	[0.65, 0.90]	[50, inf]
Heavy	[0, 0.65]	
Heavy*	[0.25, 0.65]	
All	[0, 0.90]	

C. Implementation Details

Anchor Generator. In our experiment, the sizes of the anchors are set to 8N, where N=4, 8, 16, 32, 64, and N correspond to the downsampling factor of the FPN five-layer feature maps. And the anchors in each layer are set with three size ratios of 1:1, 1:2, and 2:1.

Loss Function. We use the loss function L to optimize the overall network:

$$L = \lambda_1 L_0 + \lambda_2 L_{mask} \tag{12}$$

Where $\lambda_1 = \lambda_2 = 1$; L_0 is defined the same as (1) and L_{mask} is defined the same as [41].

Optimizer. All experiments are performed on the NVIDIA 3090 GPU. we finetune the ResNet-50 model which has been pre-trained on the ImageNet dataset. During training, each iteration processes 2 images with a total of 20 epochs. We use the SGD optimizer with 0.9 momentum and 0.0001 weight decay to optimize our model. In the optimization process, first, we use the linear warm-up strategy to warm up the learning rate from 0.0025×10^{-3} to 0.0025 in the first 500 iterations, and then keep it unchanged and train to the 15th epoch. From the 16th epoch, the learning rate decays with a fixed step length until the end of the training.

D. Ablation Study

In this part, we carry out the experiments on the CityPersons and CrowdHuman validation set, which include the comparison between each module of CCFA-Net and the baseline method, and the comparison of three attention mechanism fusion methods.

Baseline Comparison. we compare each module of the network with the baseline network Faster RCNN[26] (FR), and the results are shown in Table II . The MR^{-2} of FR on the Partial, Heavy, and All subsets of CityPersons are 15.59%, 53.99%, and 40.95%, respectively, whereas the CCFA-Net achieves 15.42%, 52.87%, and 39.87% MR^{-2} on these three subsets, which outperforms FR by 0.17%, 1.12%, and 1.14%. The results prove the effectiveness of our proposed fusion attention mechanism. Compared with the performance on the Partial subset, the MR^{-2} of the fusion attention mechanism in the Heavy subset is reduced by 1.14%., which illustrates that the CCFA-Net has a more effective improvement in the case of extremely severe occlusion.

In addition, we compare the results of correlation attention and correction attention acting independently on the network with the baseline. The correction attention achieves an absolute reduction of 0.71% and 0.7% on Heavy and All subsets, and the correlation attention achieves 1.75% and 0.88%, respectively, which demonstrates their effectiveness. which demonstrates its effectiveness. Comparing the improvement effect of two independent attention and fusion attention, it can be seen that although the two independent attention mechanisms achieve a reduction of MR^{-2} on the Heavy and All subsets, they perform worse on the Partial subset, whereas the fusion attention mechanism obtains a gain on all the subsets. These results show that a single attention mechanism can improve the overall occlusion situation, but it cannot adapt to the multi-pattern scenes, whereas the fusion attention mechanism can overcome the shortcomings of

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3242082

independent attention thereby improving the scene adaptability and overall performance of the network.

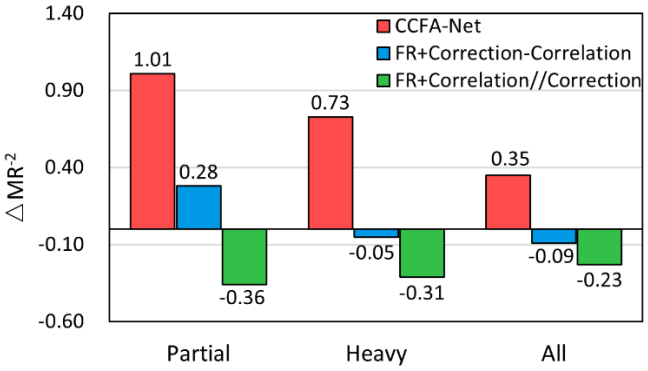


Fig. 9. Test results of three different fusion structures of the attention module on the CrowdHuman validation subset. The origin of the coordinate axis is the MR^{-2} of the baseline. The data shows the improvement of each algorithm compared with the baseline.

Fusion Methods Comparison. we compare three ways of fusion methods of attention modules, namely 'correlation-correction' serial connection, 'correction-correlation' serial connection, and 'correlation//correction' parallel connection. The results on CityPersons datasets are shown in Table III. The two serial connection methods on the All subsets perform better than the baseline FR, where the 'correction-correlation' fusion method achieves 0.34% improvement, and the 'correlation-correction' achieves 1.14%, whereas the parallel connection drops 0.18% compared to FR. Also, we conduct the experiments on CrowdHuman datasets as shown in Fig. 9. Our proposed method of ' correlation-correction' serial connection achieves 1.01%, 0.73% and 0.35% improvement in three subsets than the baseline method. However, the 'correction-correlation' serial connection method only works in the Partial subset, and the 'correlation//correction' parallel connection even drops 0.36%, 0.31% and 0.23% compared to the baseline.

The results indicate that the fusion method of the serial connection of the two attentions can improve the overall occlusion situation, whereas the parallel connection leads to a poor effect on the network. In addition, the 'correlation-correction' serial model performs better than the 'correction-correlation' one, which verifies that the serial connection method of "correlation first, correction behind" is the most effective attention mechanism fusion method in the scene we studied.

TABLE II

THE TEST RESULTS OF THE DETECTORS WITH DIFFERENT ATTENTION MODULES ON THE CITYPERSONS VALIDATION SET. THE BASELINE METHOD IS FASTER RCNN. ΔMR^{-2} REPRESENTS THE PERFORMANCE GAIN OVER THE BASELINE METHOD. BOLD FONTS INDICATE RESULTS THAT OUTPERFORM THE BASELINE METHOD.

Detector	Partial		Heavy		All	
	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$
FR	15.59	-	53.99	-	40.95	-
FR+Correlation	16.02	-0.43	52.24	**+1.75**	40.07	**+0.88**
FR+Correction	15.62	-0.03	53.28	**+0.71**	40.25	**+0.7**
CCFA-Net	15.42	**+0.17**	52.87	**+1.12**	39.87	**+1.14**

TABLE III

TEST RESULTS OF THREE DIFFERENT FUSION STRUCTURES OF THE ATTENTION MODULE ON THE CITYPERSONS VALIDATION SUBSET.

Detector	Partial		Heavy		All	
	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$
FR	15.59	-	53.99	-	40.95	-
FR+Correlation//Correction	16.04	-0.45	54.45	-0.46	41.13	-0.18
FR+Correction-Correlation	15.80	-0.21	54.32	-0.33	40.61	**+0.34**
FR+Correlation-Correction (CCFA-Net)	15.42	**+0.17**	52.87	**+1.12**	39.87	**+1.14**

E. Comparison with the State-of-the-art

CityPersons. We compare the proposed method with the advanced methods on the CityPersons dataset, namely Repulsion Loss[44], OR-CNN[45], ATT-vbb[18], ATT-part[18], MGAN[13], PODE+RPN[12], Pedhunter[14], TLL[46], MSAGNet[47], FRCN+A+PT[48] and FRCN+A+DT[48]. Since the above methods have different definitions for the visibility range of different occlusion modes, we will compare with them on the Partial, Heavy, and Heavy* subsets to be consistent with the above methods. In addition, since the size of the input image has a greater impact on the test results, we will test the model on the original image size and the size of the original image enlarged by 1.3 times. The results on the Partial and Heavy subsets in Table IV show that when the original image size is enlarged by 1.3 times our method obtains an MR^{-2} of 13.35% and 49.99%, respectively, which outperforms all the methods compared. Table V shows the comparisons on the Heavy* subset. When the input is the original image size, our method achieves a MR^{-2} of 45.49%, which is 6.21% lower than the best-performing method MSAGNet. When the original image size is enlarged by 1.3 times, our method obtains a MR^{-2} of 42.21%, which achieves 1.32% improvement over the advanced method Pedhunter.

TABLE IV

COMPARISON OF TEST RESULTS OF CCFA-NET AND OTHER STATE-OF-THE-ART METHODS ON THE PARTIAL AND HEAVY SUBSETS OF THE CITYPERSONS VALIDATION SET. THE NUMBERS IN THE TABLE REPRESENT MR^{-2}, THE BOLD FONT REPRESENTS THE OPTIMAL RESULT.

Detector	Scale	Partial	Heavy
Repulsion Loss	1×	16.8	56.9
TLL	1×	17.2	53.6
CCFA-Net	1×	**15.42**	**52.88**
Repulsion Loss	1.3×	14.8	55.3
OR-CNN	1.3×	13.7	51.3
CCFA-Net	1.3×	**13.35**	**49.99**

TABLE V
COMPARISON OF TEST RESULTS OF CCFA-NET AND OTHER STATE-OF-THE-ART METHODS ON THE HEAVY* SUBSET OF THE CITYPERSONS VALIDATION SET.

Detector	Scale	Heavy*
ATT-vbb	1×	57.31
ATT-part	1×	56.66
MGAN	1×	51.7
MSAGNet	1×	49.3
CCFA-Net	1×	**45.49**
PODE+RPN	1.3×	44.15
FRCN+A+PT	1.3×	45.8
FRCN+A+DT	1.3×	44.3
Pedhunter	1.3×	43.53
CCFA-Net	1.3×	**42.21**

Caltech. To test the generalization ability of the model on different datasets, we compare the proposed method on the Caltech-USA dataset with multiple methods on Heavy, Partial and All subsets, namely SCCPriors[49], UDN+[50], RPN+BF[51], Checkerboards+[52], ShearFtrs[53], ACF++[54], LDCF++[54], TACNN[55], MRFC+Semantic[56]. As is shown in Fig. 10, on the Partial and Heavy subsets, our method achieves the MR^{-2} of 20.40% and 54.33%, which are better than all comparison methods. On the All subset, our method reaches 58.79%, which obtains 1.09% improvement over the advanced method RPN+BF.

In addition, since the visibility of pedestrians in most images in the Caltech dataset is between [0.65, 1], there are very few cases of severe occlusion. Therefore, we add the experiments in the Reasonable subsets whose visibility is in the range of [0.65, 1]. We compare our methods with some advanced methods, namely UDN+[50], PCN[57], CompACT-Deep[58], FasterRCNN+ATT[18], SA-FastRCNN[59], RPN+BF[51], F-DNN2+SS[60], MS-CNN[33], DeepParts[10] and GDFL[61]. It can be seen from Table VI that our method is superior to all compared methods, reaching a MR^{-2} of 5.66%, which obtains 0.26% improvement over the advanced method GDFL[61].

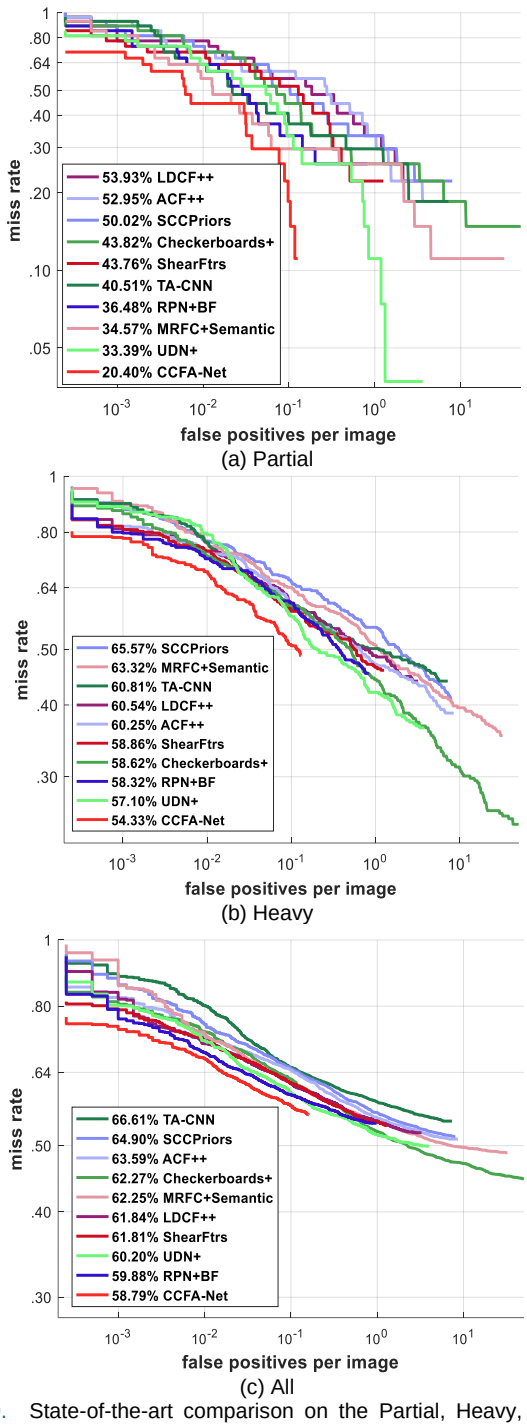


Fig. 10. State-of-the-art comparison on the Partial, Heavy, and All subsets of the Caltech dataset. The horizontal axis is FPPI, and the vertical axis is the missed detection rate. CCFA-Net outperforms other methods on all three subsets.

F. Extension to Multi-Classification

Pedestrians in complex traffic scenes have different states, including cyclists, people walking, people in wheelchairs, and so on. There are significant differences in the characteristics of people in various states, resulting in more complex occlusion patterns. To improve the expressiveness of the model in this type of complex scene, we retain the multi-class features by generating multiple masks in the correction attention module, so that it can accurately detect multi-class pedestrians under complex occlusion modes.

CityPersons. In this part of the experiment, we use the CityPersons dataset to verify the effectiveness of our method for multi-class detection. The four types of pedestrian annotations in the CityPersons dataset were used during training and testing, including 'pedestrian', 'rider', 'sitting person', and 'other person'. The comparison results between our method and the baseline method Faster RCNN are shown in Table VII. It can be seen from the data that our method achieves the MR^{-2} of 15.27%, 50.28%, and 38.44% on the Partial, Heavy, and All subsets, respectively. Compared with Faster RCNN, it reduces MR^{-2} by 0.71% in the All subset, and the performance on the Partial subset is significantly improved, where the MR^{-2} drops 1.02%. It can be seen that the fusion attention mechanism we proposed still has excellent performance on multi-class pedestrian detection tasks.

It is worth noting that most of the current researches on pedestrian detection in occluded scenes only focus on single-class pedestrians, whereas our proposed fusion attention mechanism has achieved outstanding performance in multi-class pedestrian detection scenes, which provides new ideas for subsequent pedestrian detection studies in occluded scenes.

Qualitative Test. In the CityPersons dataset, 83% of the objects are "pedestrian", and the other three types of pedestrians account for less. Therefore, to more effectively test the performance of our method to detect multi-class pedestrians in complex occluded traffic environments, we collected images that meet the requirements of the scene in our research and employ them in qualitative testing of the model. In this part, the multi-class CCFA-Net trained on the CityPersons training set will be tested in our data, and the results are shown in Fig. 11. From the comparison in the figure, it can be seen that, in multi-class pedestrian detection, due to the participation of vehicles such as bicycles and motorcycles, the occlusion of pedestrians will be more complicated than that of a single type of pedestrian, so it is tough for a general object detection network to overcome a large number of false detection. The detection result of Faster RCNN in Fig. 11(b) shows its difficulty in the case of severe occlusion, whereas in Fig. 11(c) our CCFA-Net shows obvious alleviation of the above problems. The excellent results not only show the effectiveness of our proposed model for multi-class pedestrian detection in complex occlusion scenes but also show that the method has excellent scene generalization ability.

TABLE VI
STATE-OF-THE-ART COMPARISON ON THE REASONABLE SUBSET OF THE CALTECH DATASET. CCFA-NET OUTPERFORMS OTHER METHODS ON THE REASONABLE SUBSET.

Detector	Reasonable
DeepParts	10.62
UDN+	8.34
MS-CNN	7.84
PCN	7.58
CompACT-Deep	7.56
FasterRCNN+ATT	7.52
SA-FastRCNN	6.66
RPN+BF	6.49
F-DNN2+SS	6.23
GDFL	5.92
CCFA-Net	**5.66**

(a) Input

(b) Faster RCNN

(c) CCFA-Net

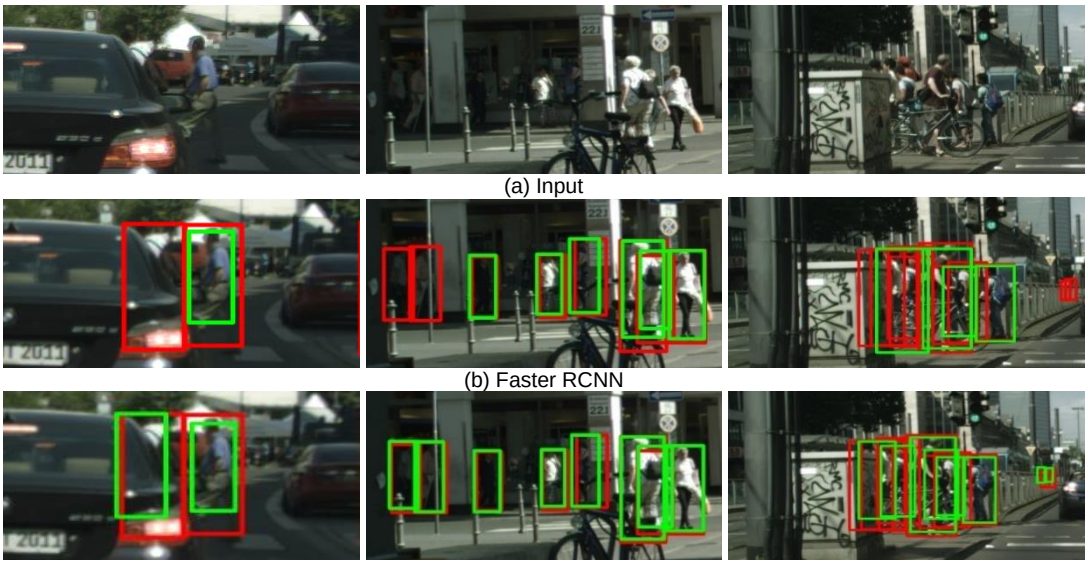
Fig. 11. Qualitative test results of multi-class pedestrian detection in crowded scenes. (a) shows the original image we collected; (b) shows the detection result of Faster RCNN; (c) shows the detection result of CCFA-Net. When detecting heavily occluded multi-class pedestrians, CCFA-Net effectively improves the false and missed detections of Faster RCNN.

TABLE VII
TEST RESULTS OF CCFA-NET AND BASELINE METHOD FASTER RCNN FOR MULTI-CLASS PEDESTRIAN DETECTION ON CITYPERSONS VALIDATION SET.

Detector	Partial		Heavy		All	
	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$	MR^{-2}	$\triangle MR^{-2}\downarrow$
Faster RCNN	16.29	-	50.36	-	39.15	-
CCFA-Net	15.27	**+1.02**	50.28	**+0.08**	38.44	**+0.71**

V. CONCLUSION

In this paper, a correlation-and-correction fusion attention mechanism based pedestrian detection network is proposed to solve the pedestrian detection problem in occluded scenes. Specifically, a multi-mask correction attention module is proposed to strengthen the visible part features of occluded pedestrians and correct the false one. And a correlation attention module is fused with the module to globally enhance the correlation of pedestrians' body part features. What's more, the effective fusion methods of the two modules are studied. Besides, we extend CCFA-Net to the multi-class pedestrian detection in occluded scenes. Extensive experimental results on the CityPersons Caltech, and CrowdHuman datasets demonstrate the effectiveness and advancement of our method.

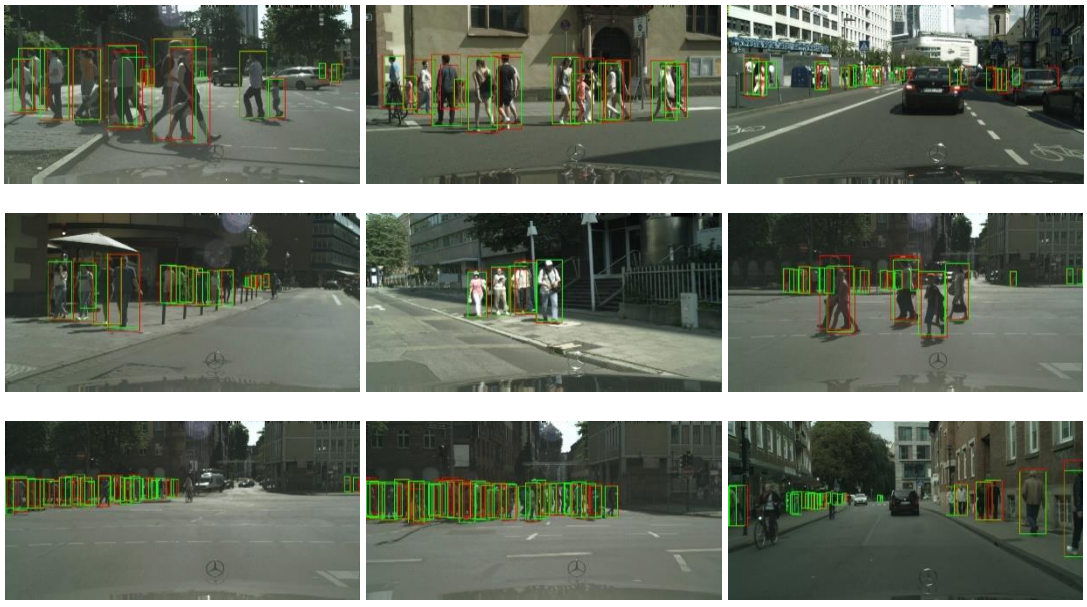


(a) Input

(b) Faster RCNN

(c) CCFA-Net

Fig. 12. Qualitative comparison results on the CityPersons validation set. (a) shows the input image, (b) shows the detection result of Faster RCNN, and (c) shows the detection result of CCFA-Net. The red boxes are the ground truth of the targets, and the green boxes are the detection results of each method.



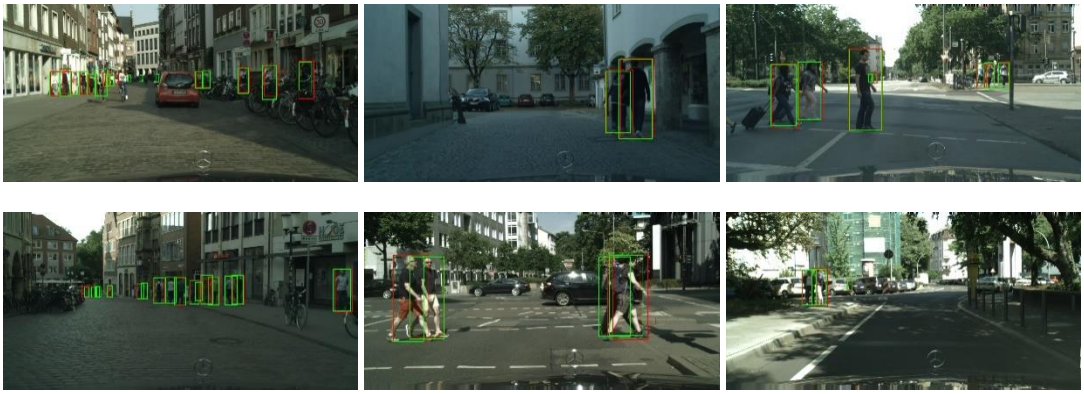


Fig. 13. Qualitative detection results on the CityPersons validation set. The red boxes are the ground truth, and the green boxes are the detection results of CCFA-Net.

REFERENCES

[1] D. K. Dewangan and S. P. Sahu, "Driving behavior analysis of intelligent vehicle system for lane detection using vision-sensor," *IEEE Sensors J.,* vol. 21, no. 5, pp. 6367-6375, Mar. 2021.

[2] G. Li, S. Lin, S. Li, and X. Qu, "Learning Automated Driving in Complex Intersection Scenarios Based on Camera Sensors: A Deep Reinforcement Learning Approach," *IEEE Sensors J.,* vol. 22, no. 5, pp. 4687-4696, Mar. 2022.

[3] R. Zhang *et al.*, "A Multi-Vehicle Longitudinal Trajectory Collision Avoidance Strategy Using AEBS With Vehicle-Infrastructure Communication," *IEEE Trans. Veh. Technol.,* vol. 71, no. 2, pp. 1253-1266, Feb. 2021.

[4] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 3213-3221.

[5] 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 Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1259-1267.

[6] H. Luo, P. Wang, H. Chen, and M. Xu, "Object detection method based on shallow feature fusion and semantic information enhancement," *IEEE Sensors J.,* vol. 21, no. 19, pp. 21839-21851, Oct. 2021.

[7] F. Altay and S. Velipasalar, "The Use of Thermal Cameras for Pedestrian Detection," *IEEE Sensors J.,* vol. 22, no. 12, pp. 11489 - 11498, Jun. 2022.

[8] W. Wei, L. Cheng, Y. Xia, P. Zhang, J. Gu, and X. Liu, "Occluded pedestrian detection based on depth vision significance in biomimetic binocular," *IEEE Sensors J.,* vol. 19, no. 23, pp. 11469-11474, Dec. 2019.

[9] J. Xie, Y. Pang, H. Cholakkal, R. Anwer, F. Khan, and L. Shao, "PSC-Net: learning part spatial co-occurrence for occluded pedestrian detection," *Sci. CHINA Inform. Sci.,* vol. 64, no. 2, pp. 1-13, Nov. 2021.

[10] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Santiago, Chile, 2015, pp. 1904-1912.

[11] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: a survey," *IEEE Trans. Pattern Anal.,* Apr. 2021.

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

[13] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seoul, Korea (South), 2019, pp. 4967-4975.

[14] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 10639-10646.

[15] P. Zhou *et al.*, "Noh-nms: Improving pedestrian detection by nearby objects hallucination," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1967-1975.

[16] Z. Zhang, Y. Wang, H. Jiang, and X. Zeng, "Strict NMS: Pedestrian Detection in Crowd Scenes," in *Proc. IEEE 3rd Int. Conf. Inf. Syst. Comput Aided Educ. (ICISCAE)*, Dalian, China, 2020, pp. 225-230.

[17] 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)*, Long Beach, CA, USA, 2019, pp. 6459-6468.

[18] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6995-7003.

[19] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal.,* vol. 34, no. 4, pp. 743-761, Apr. 2011.

[20] S. Shao *et al.* "Crowdhuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*. [Online]. Available: https://arxiv.org/abs/1805.00123

[21] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5363-5371.

[22] Z. Yi, S. Yongliang, and Z. Jun, "An improved tiny-yolov3 pedestrian detection algorithm," *Optik,* vol. 183, pp. 17-23, Apr. 2019.

[23] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 5187-5196.

[24] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 11346-11355.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 580-587.

[26] 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. 2017.

[27] X. Song, S. Gao, and C. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alex. Eng. J.,* vol. 60, no. 1, pp. 73-85, 2021.

[28] G. Shen, L. Zhu, J. Lou, S. Shen, Z. Liu, and L. Tang, "Infrared multi-pedestrian tracking in vertical view via siamese convolution network," *IEEE Access,* vol. 7, pp. 42718-42725, Jan. 2019.

[29] X. Dai *et al.*, "Multi-task faster R-CNN for nighttime pedestrian detection and distance estimation," *Infrared Phys. Techn.,* vol. 115, p. 103694, Jun. 2021.

[30] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3213-3221.

[31] Y. Jin, Y. Zhang, Y. Cen, Y. Li, V. Mladenovic, and V. Voronin, "Pedestrian detection with super-resolution reconstruction for low-quality image," *Pattern Recogn.,* vol. 115, p. 107846, Jul. 2021.

[32] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4681-4690.

[33] 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. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 354-370.

[34] C. Zhou and J. Yuan, "Multi-label learning of part detectors for occluded pedestrian detection," *Pattern Recogn.,* vol. 86, pp. 99-111, Feb. 2019.

[35] 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)*, Seattle, WA, USA, 2020, pp. 10750-10759.

[36] Q. Wang, B. Wu, P. Zhu, P. Li, and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020.

[37] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 12894-12904.

[38] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan. "Volo: Vision outlooker for visual recognition," 2021, *arXiv:2106.13112*. [Online]. Available: https://arxiv.org/abs/2106.13112

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal.,* vol. 42, no. 8, pp. 2011-2023, Aug. 2020.

[40] X. Li *et al.*, "Weighted feature pyramid networks for object detection," in *Proc. - 2019 IEEE Intl Conf Parallel Distrib. Process. with Appl. Big Data Cloud Comput. Sustain. Comput. Commun. Soc. Comput. Networking (ISPA/BDCloud/SocialCom/SustainCom)*, Xiamen, China, 2019, pp. 1500-1504.

[41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 2961-2969.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 618-626.

[43] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 3213-3223.

[44] 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 Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7774-7783.

[45] 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. (ECCV)*, Munich, Germany, 2018, pp. 637-653.

[46] 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. (ECCV)*, Munich, Germany, 2018, pp. 536-551.

[47] H. Zhang, C. Yan, X. Li, Y. Yang, and D. Yuan, "MSAGNet: Multi-Stream Attribute-Guided Network for Occluded Pedestrian Detection," *IEEE Signal Process Lett.,* vol. 29, pp. 2163-2167, Oct. 2022.

[48] Z. Chunluan, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea, 2019, pp. 9557-9566.

[49] Y. Yang, Z. Wang, and F. Wu, "Exploring Prior Knowledge for Pedestrian Detection," in *Proc. British Mach. Vis. Conf*, 2015, pp. 176.1-176.12.

[50] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Trans. Pattern Anal.,* vol. 40, no. 8, pp. 1874-1887, Aug. 2017.

[51] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 2016, pp. 443-457.

[52] Y. Kuranuki and I. Patras, "Minimal filtered channel features for pedestrian detection," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, 2016, pp. 681-686.

[53] L. Pfeifer and Vision, "Shearlet features for pedestrian detection," *J. Math. Imaging,* vol. 61, no. 3, pp. 292-309, Jul. 2019.

[54] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? on the limits of boosted trees for object detection," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, 2016, pp. 3350-3355.

[55] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 5079-5087.

[56] A. D. Costea and S. Nedevschi, "Semantic channels for fast pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2360-2368.

[57] S. Wang, J. Cheng, H. Liu, and M. Tang. "Pcn: Part and context information for pedestrian detection with cnns," 2018, *arXiv:1804.04483*. [Online]. Available: https://arxiv.org/abs/1804.04483

[58] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for pedestrian detection," *IEEE Trans. Pattern Anal.,* vol. 42, no. 9, pp. 2195-2211, Sep. 2019.

[59] 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, Apr. 2017.

[60] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. Davis. "Fused deep neural networks for efficient pedestrian detection," 2018, *arXiv:1805.08688*. [Online]. Available: https://arxiv.org/abs/1805.08688

[61] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 732-747.

Fengmin Zou received the B.S. degree in Information Science and Engineering from Northeast University, Shenyang, China, in 2020. She is currently pursuing the M.S. degree in Instrument Science and Technology with Southeast University, Nanjing, China.

Her current research interests include deep learning, computer vision, object detection and their application on autonomous driving.

Xu Li (Member, IEEE) received the Ph.D. degree in instrument science and technology from Southeast University, Nanjing, China, in 2006.

He is currently a Professor with the School of Instrument Science and Engineering, Southeast University. His current research interests include the collaborative perception and control of intelligent vehicle and infrastructure systems, information fusion, automated vehicles, and active safety.

Qimin Xu received the B.S., M.S., and Ph.D. degrees in instrument science and technology from Southeast University, Nanjing, China, in 2011, 2014, and 2018, respectively.

He is currently a Lecturer with the School of Instrument Science and Engineering, Southeast University. His current research interests include vehicle state estimation, vehicle positioning, and autonomous driving.

Zhengliang Sun received the B.S. degree from Southeast University, Nanjing, China, in 1987 and the M.S. degree in software engineering from Peking University, Beijing, China, in 2011.

He is currently a Research Fellow with the Traffic Management Research Institute, Ministry of Public Security, Beijing, China. His current re- search interests include traffic information man- agement and intelligent transportation system.

Jianxiao Zhu is a PhD student at Southeast University, Nanjing, China.

His research interests in the smart transportation system and machine learning. His current research is the application of multi-sensor fusion algorithms in the smart vehicle-infrastructure cooperative system.