IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 April 2024, accepted 20 May 2024, date of publication 29 May 2024, date of current version 5 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406968

RESEARCH ARTICLE

Multi-Scale Structure Perception and Global Context-Aware Method for Small-Scale Pedestrian Detection

HAO GAO[1], SHUCHENG HUANG[1], MINGXING LI[2], AND TIAN LI[3]
[1]School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China
[2]Jingjiang College, Jiangsu University, Zhenjiang 212013, China
[3]Suzhou Institute of Technology, Jiangsu University of Science and Technology, Suzhou 215699, China

Corresponding author: Tian Li (tli@just.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62276118 and Grant 61772244.

ABSTRACT In pedestrian detection, small-scale pedestrians often face challenges such as limited pixel values and insufficient features, often leading to wrong or missed detection. Therefore, this paper proposed a multi-scale structure perception and global context-aware method for small-scale pedestrian detection. Firstly, to address the issue of decreasing features caused by the network deepens, we designed a feature fusion strategy to overcome the constraints of the feature pyramid hierarchy. This strategy combines deep and shallow feature maps and leverages the advantages of Transformer to capture long-distance dependent features, incorporating a global context information module to retain a substantial amount of small-scale pedestrian features. Secondly, considering the confusion between small-scale pedestrian features and background information, we employed a combination of self-attention modules and channel attention modules to jointly model the spatial and channel correlations of feature maps. This utilization of small-scale pedestrian context and channel information enhances small-scale pedestrian features while suppressing background information. Finally, to address the issue of gradient explosion during model training, we introduced a novel weighted loss function named ES-IoU, which significantly improved the convergence speed. Extensive experimental results on the CityPersons and CrowdHuman datasets demonstrate that the proposed method achieves a substantial improvement upon state-of-the-art methods.

INDEX TERMS Context information, self-attention, small-scale pedestrian detection, Transformer.

I. INTRODUCTION

Pedestrian detection, as a specialized branch of general object detection, has been widely studied and applied in both academia and industry. In recent years, significant progress has been made in pedestrian detection through the successful application of deep convolutional neural networks [1], [2], [3], [4], [5], [6], [7], [8], [9]. But in complex practical scenes, especially for small-scale pedestrians, these current methods suffer from wrong or missed detection due to confusing human-like objects or heavily occlusion.

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

Anchor-based and anchor-free pedestrian detection are two major approaches based on convolutional neural networks. Typical anchor-based methods include Faster R-CNN and its derivatives [10], [11], [12]. These methods generate candidate proposals beforehand and then use a classifier to determine whether each proposal contains a pedestrian. Anchor-based models are time-consuming, and most candidate proposals provide limited information. To address these issues, some researchers have proposed anchor-free detectors that can directly predict pedestrians from images [13]. In other words, redundant steps such as defining anchors and extracting features from candidate regions are skipped, and pedestrians are predicted directly from the raw images. For example, ALFNet adopts a progressive localization fitting strategy to

continuously optimize default anchor boxes [14]. For anchor-free detectors, CSP can locate pedestrian targets by directly predicting the center point, width, and height of bounding boxes [15].

Although the above methods have achieved good results in pedestrian detection, but they are mainly used for general pedestrian detection and perform poorly in detecting small-scale pedestrians. Based on the characteristics of small-scale pedestrian, some researchers have proposed a series of detection methods for small-scale pedestrian. Regarding the detection of small-scale pedestrians, current research mainly falls into image pyramid methods, feature pyramid methods, and other methods. Firstly, the most common image pyramid methods are Gaussian pyramids and Laplacian pyramids. Secondly, feature pyramid methods utilize feature maps of different resolutions generated by multi-layer convolutional layers for detection. Liu et al. [16] proposed the Single Shot Multibox Detector (SSD), which utilizes shallow high-resolution feature maps to detect small objects and deep low-resolution feature maps to detect large objects. Lin et al. [17] introduced Feature Pyramid Networks (FPN), which adopt a top-down approach to up-sampling deep feature maps and fuse them with the next feature map to ensure that each layer has appropriate resolution and semantic information. Finally, besides image and feature pyramid methods, researchers have proposed some novel detection approaches [18], [19].

Although existing methods can locate pedestrians in given images, their detection accuracy for small-scale pedestrians is low and prone to missed detections, which is one of the core challenges in pedestrian detection.

To address the detection problem of small-scale pedestrians, this paper constructs a detector based on a multi-scale structure perception and global contextual information. The method enhances the features of small-scale pedestrians by focusing on their characteristics in convolutional neural networks to improve the accuracy of small-scale pedestrian detection. The main contributions of this paper can be summarized as follows:

(1) We designed a feature fusion strategy to break through the constraints of the hierarchical structure of the feature pyramid, by integrating deep and shallow feature maps. Additionally, leveraging the advantages of Transformer to capture long-distance dependent features, a global contextual information module is designed to retain a large amount of small-scale pedestrian features.
(2) We employed a combination of self-attention modules and channel attention modules to model the spatial and channel correlations of feature maps. By incorporating the contextual and channel information of small-scale pedestrians, this approach enhances their features while suppressing background information.
(3) We proposed a new weighted loss function, ES-IoU, which can alleviate the gradient explosion phenomenon and effectively improve the convergence speed of the network.

II. RELATED WORKS

A. PEDESTRIAN DETECTION

Currently, computer vision based on deep learning technology is rapidly expanding, many variants of Faster R-CNN [10], such as SA-Fast RCNN [20] and MS-CNN [21], have achieved improved detection performance by directly solving the problem of target scale. Although two-step detectors are widely used, they mostly use visual information only to locate pedestrian objects in images. Cascade R-CNN [8] is a multi-step detection model.

This method continuously increases the Intersection over Union (IoU) threshold, enabling the model to better regress on the generated proposals, ultimately training a high-quality detector. Wang proposed Repulsion Loss [48] to improve pedestrian detection accuracy in occluded scene. Repulsion loss includes three parts: the loss value between the predicted box and the target ground-truth box; loss value between the predicted box and adjacent target ground-truth box; loss value between predicted boxes and adjacent predicted boxes that are not predicting the same real target. Xie et al. proposed the MGAN network [22], the network emphasized on visible pedestrian regions while suppressing the occluded ones by modulating full body features. Xu et al. [52] designed a joint prediction scheme, which is executed through an assignment of bounding boxes and a joint loss to improve the accuracy of pedestrian detection.

B. SMALL-SCALE PEDESTRIAN DETECTION

In recent years, researchers have devoted significant efforts to overcoming challenges in small-scale pedestrian detection. Song et al. [13] proposed a pedestrian detection network which is based on vertical lines by using the vertical characteristics of upright pedestrians. This approach does not need to set additional prior box parameters, but directly use feature maps for classification prediction and location regression. Ding et al. [40] proposed a learnable Dynamic HRNet (DHRNet) to generate different network paths adaptive to different scales. Xie and Wang [23] design a feature enrichment unit to produce more representative features to improve small-scaled pedestrian detection performance. GDFL [24] encodes fine-grained attention masks into convolutional feature maps, which enables the model to pay more attention to small-scale pedestrian information. Li et al. [25] proposed a novel perceptual generative adversarial network, which narrows the representation gap between small and large targets, making the characteristics of small targets closer to those of large targets, ultimately improving the detection of small targets. Hu et al. [26] designed attention mechanism weights to utilize the interrelationship between objects in images, providing more surrounding information for small objects to aid in recognition and thus improving detection accuracy. Tan et al. [27] proposed a Bidirectional Feature

Enhancement Module, which enhances the semantic information of low-level features and enriches the localization information of high-level features.

C. REGRESSION LOSS FUNCTION BASED ON IOU

The regression loss function based on Intersection over Union (IoU) is one of the key techniques for evaluating the similarity between predicted bounding boxes and ground truth bounding boxes in object detection tasks. Yu et al. proposed a method called Unitbox [28]. This method achieved more effective results by directly using IoU as a regression loss function for object detection. However, IoU fails to reflect the distance between bounding boxes when they do not intersect. Researchers proposed a generalized IoU with scale invariance and introduced the maximum enclosing rectangle of two boxes to improve IoU calculation [29]. However, subsequent studies revealed slow convergence issues with generalized IoU. To address this problem, DIoU [30] introduced a new metric that considers geometric measures such as overlap area, aspect ratio, and center distance between boxes, enhancing object detection performance. CIoU [30] loss further improved upon DIoU by considering additional geometric factors such as overlap area, aspect ratio, and center distance, resulting in improved accuracy and stability of object detection algorithms. However, it is unfair for CIoU to evaluate predictions of pedestrians at all scales using the same loss, as pedestrians of different scales occupy different proportions of pixels during training. To address this issue, we propose a loss function, ES-IoU, more suitable for small-scale pedestrian detection. The improved loss function not only enhances model accuracy in small-scale pedestrian detection but also resolves gradient explosion issues caused by excessively small targets.

III. METHODOLOGY

To enable the network to integrate features of different scales while maintaining focus on the target itself, this paper proposed a method based on multi-scale structural perception for small-scale pedestrian detection. Additionally, to better utilize global information, a global contextual information module is designed. Considering the problem of small-scale pedestrian features being easily confused with background information, a feature enhancement module is constructed by combining self-attention modules and channel attention modules to model feature map spatial and channel correlations. This module utilizes both the context information and channel information of small-scale pedestrians to enhance their features while suppressing background information. Finally, a loss function more suitable for small-scale pedestrian detection, ES-IoU, is proposed, which can improve both the convergence speed of model training and detection accuracy.

A. NETWORK ARCHITECTURE

As shown in Figure 1, the model consists of three key components. For input images, YOLOv5 is chosen as the baseline network to extract pedestrian features, obtaining multi-layer feature maps with varying resolutions from shallow to deep layers. YOLOv5 balances detection accuracy and real-time performance, not only meeting the needs of real-time image object detection but also having a smaller structure. YOLOv5 is relatively mature in pedestrian detection. Hence, we used YOLOv5 as the basic pedestrian detection model. Subsequently, a feature pyramid module is employed to fuse shallow and deep feature maps, facilitating the flow of information between high-level and low-level features. To obtain more comprehensive structural feature information, particularly for small-scale objects with fewer pixel points, this paper designs a Transformer-based Global Context Information Module (TGCM) to further enrich the semantic information in the deeper layers, compensating for the loss of details in small targets in deep layers. Then, a feature enhancement module is utilized to enhance the features of small-scale pedestrians, guiding the network to focus on small-scale pedestrians. Finally, the detection module completes the prediction of classification, regression, and position information for feature points, thereby obtaining the predicted bounding boxes.

B. FEATURE PYRAMID MODULE

This paper utilized a bidirectional feature pyramid to facilitate the information flow of $\{F_2, F_3, F_4, F_5\}$ feature maps. Among them, the shallow feature map $\{F_2\}$ with high resolution can provide more accurate position information and edge shapes, while the deep feature map $\{F_5\}$ with lower resolution possesses stronger semantic information. The bidirectional feature pyramid up-samples the deep features to increase resolution and integrates them with shallow feature maps, and then down-samples the shallow feature maps to decrease resolution and integrates them with deep feature maps. This feature pyramid effectively retains pedestrian information from shallow feature maps and compensates for the one-way information flow during the top-down feature fusion process, laying a solid foundation for the enhancement of small-scale pedestrian features in the next step.

C. GLOBAL CONTEXT INFORMATION MODULE

As is well known, the comprehensive perspective of an image holds significant contextual cues, suggesting that surrounding objects near a small target can enhance detection outcomes. Moreover, given that convolution operates locally, constrained by the convolutional kernel's size, it primarily computes correlations among neighboring pixels, thus limiting the utilization of global contextual cues. To address this limitation and capture inter-pixel relationships across various regions, this study introduces a Transformer-based Global Context Information Module (TGCM), illustrated in Figure 2, built upon the Transformer architecture [31], [32]. Given the Transformer's ability to capture extensive dependencies within image features, TGCM acts as a global mechanism, connecting the current region to others. Through learning interactions among pixels in disparate regions, TGCM

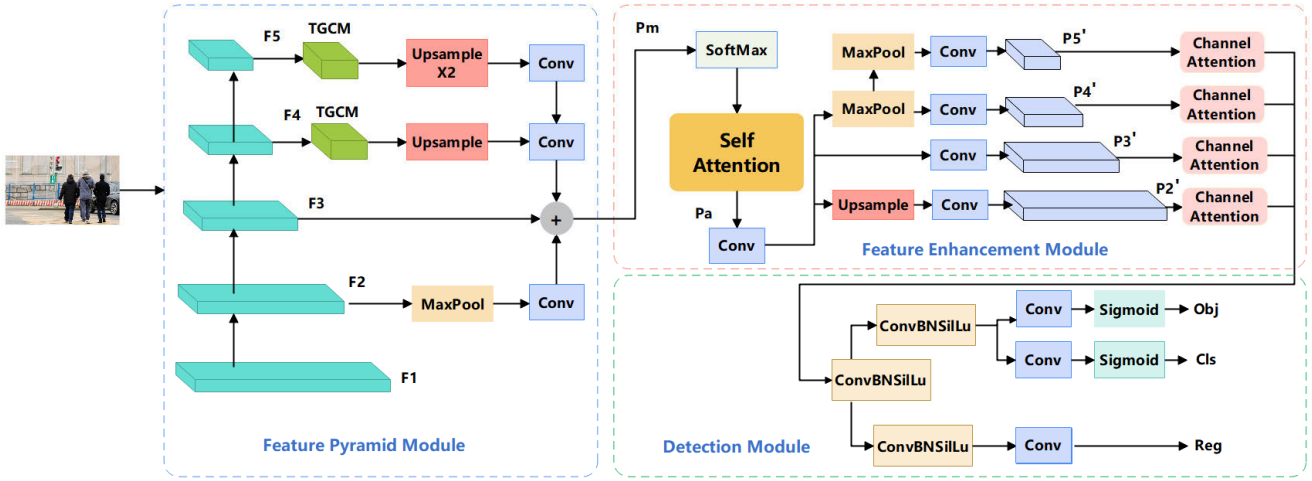

FIGURE 1. Network architecture diagram. The model consists of three modules: the Feature Pyramid Module, the Feature Enhancement Module, and the Detection Module. The Feature Pyramid Module is primarily used to merge shallow and deep feature maps, enabling the flow of high-level and low-level feature information; the Feature Enhancement Module is mainly used to enhance the features of small-scale pedestrians, guiding the network to focus on small-scale pedestrians; the Detection Module is primarily used to perform classification, regression, and prediction of positional information for feature points, resulting in predicted bounding boxes.

effectively harnesses global information, mitigating the challenge of inadequate feature representation associated with small targets and enhancing the model's detection performance in such scenarios. Notably, the ConvBN-SiLu in Figure 2 represents convolution (Conv), batch normalization (BN), and activation function (SiLu). The original Transformer incorporates LayerNorm normalization, primarily tailored for variable-length text sequences, whereas image dimensions typically remain uniform. Therefore, the LayerNorm of the original design model, represented as Transformer-Encoder in the figure, is omitted.

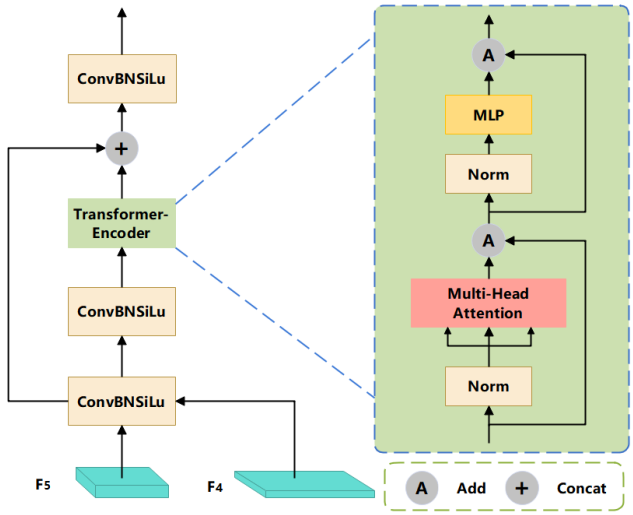

FIGURE 2. Transformer-based Global Context Information Module (TGCM).

D. FEATURE ENHANCEMENT MODULE

Based on Convolutional Neural Networks, small-scale pedestrian features exhibit two characteristics. First, they are scarce and mostly concentrated in shallow feature maps. Among the $\{C_2, C_3, C_4, C_5\}$ feature maps obtained through the feature extraction network, only $\{C_2, C_3\}$ contain some small-scale pedestrian features. Although the bidirectional feature pyramid promotes information flow between deep and shallow feature maps through top-down and bottom-up approaches, the hierarchical structure of the pyramid still greatly suppresses small-scale pedestrian features. Second, the features are not prominent and are easily submerged in background noise. Although $\{C_2, C_3\}$ feature maps contain small-scale pedestrian information, inevitably, a lot of background environmental information is also included. To enhance the detection network's focus on small-scale pedestrian information and reduce background noise, thereby enhancing the detector's capability to detect small-scale pedestrians, this paper designs a Feature Enhancement Module embedded between the feature pyramid module and the prediction network, whose module structure is shown in Figure 1. First, for the $\{P_2, P_3, P_4, P_5\}$ feature maps outputted by the feature pyramid module, the feature fusion strategy breaks through the hierarchical structure of the feature pyramid, merging shallow and deep feature maps of different resolutions with equal importance, retaining a large amount of small-scale pedestrian features in shallow feature maps. Secondly, the self-attention module explores the correlation between individual feature points in the fused feature map $\{P_m\}$ and other feature points, enhancing the contextual information of target features and suppressing noisy features at the level of individual pixels. Then, after restoring the feature map $\{P_a\}$ to the original size of $\{P_2, P_3, P_4\}$, the $\{P_2', P_3', P_4'\}$ feature maps respectively utilize the channel attention module to further model the correlation between feature map channels, guiding the network to focus on small-scale pedestrians based on the response of feature map channel importance. Finally, the three feature maps

{P_2'', P_3'', P_4''} outputted by the Feature Enhancement Module enter three detection modules with identical structures, predicting target category, regression, position, and other information in the three feature maps respectively.

The following, we will provide a detailed introduction about feature fusion strategy, self-attention module and channel attention module.

1) FEATURE FUSION STRATEGY

The feature fusion strategy scales and integrates shallow to deep feature maps, ensuring each resolution feature map receives the same information from other resolution feature maps. As depicted in Figure 1, the feature maps {P_2, P_3, P_4, P_5} outputted from the bidirectional feature pyramid are scaled to the size of {P_3} feature map through max-pooling, upsampling, and convolution operations applied to {P_2, P_4, P_5} feature maps, respectively. Subsequently, the three feature maps of identical size are added together and averaged to obtain the mixed information in the {P_m} feature map. Equation (1) represents the calculation formula for the {Pm} feature map.

$$P_m = (F_m(Conv(P_2) + P_3 + F_u(Conv(P_4)) + F_u(F_u(Conv(P_5)))) \quad (1)$$

where, $F_m(\cdot)$ represents the max-pooling operation, and $F_u(\cdot)$ is the up-sampling operation utilized to adjust the resolution of feature maps {P_2, P_4, P_5}. $Conv(\cdot)$ represents the convolution operation, employed to adjust the channel count of feature maps {P_2, P_4, P_5}.

2) SELF-ATTENTION MODULE

Convolutional Neural Networks use convolution operations to extract and integrate features, which are locally connected, thus overlooking the dependency of pedestrian detection on global information. Particularly for small-scale pedestrians, their weak representation in feature maps necessitates contextual information to help the network focus on small-scale pedestrian features and suppress background noise. The self-attention module establishes the similarity between each feature point in the {P_m} feature map and other feature points, obtaining descriptors representing the spatial correlation of the feature map, as illustrated in Figure 3.

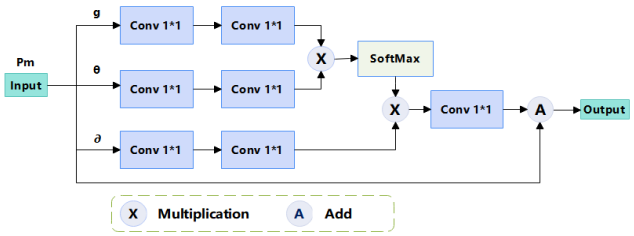

FIGURE 3. Self-attention module.

To establish the spatial correlation model of the {Pm} feature map, the self-attention module first reshapes the size of the {P_m} feature map to C × HW, representing C × HW individual feature points within {P_m} feature map. Secondly, a 1 × 1 convolution linearly maps the {P_m} feature map, yielding three separate mappings: $g(P_m)$, $\theta(P_m)$, and $\partial(P_m)$; Next, the transpose of $\theta(P_m)$, denoted as $\theta(P_m)^T \in R^{HW\times C/2}$, is multiplied by $g(P_m) \in R^{C/2\times HW}$, resulting in the feature map spatial correlation matrix $V_s \in R^{HW\times HW}$, where each value in this matrix signifies the similarity between every pair of pixels. Finally, the normalized feature map spatial correlation matrix V_s is multiplied by the original feature mapping matrix $\partial(P_m)$, yielding the self-attention response z_s for the {P_m} feature map. Equation (2) presents the calculation formula for the self-attention response z_s.

$$z_s = softmax(V_s) \cdot \partial(P_m) \quad (2)$$

$$g(P_m) = W_{g1}(W_{g2}(P_m)) \quad (3)$$

$$\theta(P_m) = W_{\theta 1}(W_{\theta 2}(P_m)) \quad (4)$$

$$\partial(P_m) = W_{\partial 1}(W_{\partial 2}(P_m)) \quad (5)$$

where W_{g1}, W_{g2}, $W_{\theta 1}$, $W_{\theta 2}$, $W_{\partial 1}$, and $W_{\partial 2}$ represent the learnable parameters in the 1 × 1 convolutional kernel respectively.

The self-attention response z_s is manifested in the form of a residual block, with Equation (6) representing the final output calculation formula of the self-attention module.

$$P_a = W_z z_s + P_m \quad (6)$$

where, W_z represents the learnable parameters in the 1 × 1 convolutional kernel.

3) CHANNEL ATTENTION MODULE

The self-attention module endows feature maps with global information in the form of attention, enabling small-scale pedestrian areas to leverage contextual information and attract the network's attention. To further enhance the features of small-scale pedestrians, inspired by the referenced paper [33], this article introduces a channel attention mechanism. the channel attention module models the correlation between feature channels to obtain descriptors that express the importance of each channel. It adaptively corrects the channel features, as illustrated in Figure 4.

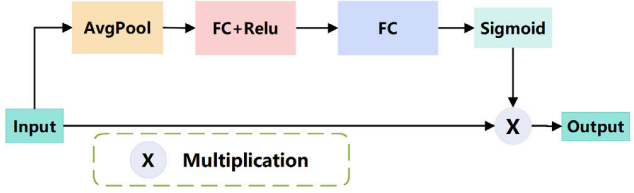

FIGURE 4. Channel attention module.

To establish the channel correlation model of the feature map, the channel attention module first compresses the global spatial information of each channel of the input feature map through average pooling, forming a feature map channel information statistical description vector $z_{ch} \in R^{C\times 1\times 1}$. Next, z_{ch} serves as the input to two consecutive fully connected

layers to predict the importance of each channel; the fully connected layer is structured with adjacent layers of neurons fully interconnected. Its function is to globally analyze z_{ch} and nonlinearly combine its channel features. In the channel attention model, the fully connected layers capture the nonlinear relationships between each channel of z_{ch}, enabling z_{ch} to adaptively adjust the description of channel importance. Finally, the sigmoid activation function outputs the importance of different channels, forming the feature channel attention vector $V_{ch} \in \mathrm{R}^{\mathrm{C}\times 1\times 1}$, where the value of each element in this vector reflects the importance of the corresponding feature channel. Equation (7) represents the calculation formula for the channel attention vector V_{ch}.

$$V_{ch} = \sigma(W_2(\delta(W_1 \cdot z_{ch}))) \tag{7}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} z_{ch} \tag{8}$$

$$\delta(x) = max(0, x) \tag{9}$$

where W_1 and W_2 are the parameters of the two fully connected layers.

The feature channel attention vector V_{ch} weights the input features channel-wise, with Equation (10) representing the final output calculation formula of the channel attention module.

$$F_{chn} = V_{ch} \otimes F \tag{10}$$

where, $\otimes$ denotes element-wise multiplication.

E. DETECTION MODULE

After the feature enhancement module, the model inputs three different resolution feature maps {P$_2$'', P$_3$'', P$_4$'', P$_5$''} into the detection module to obtain detection results. The prediction network consists of three components: category, regression, and position. Position and category predictions are simplified into a binary classification problem, using cross-entropy loss as the loss function. Equations (11) and (12) represent the calculation formulas for the loss functions of the category and position components, respectively.

$$L_{cls} = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[gt_i \cdot \lg(p_i) + (1-gt_i)\lg(1-p_i)] \tag{11}$$

where, p_i represents the predicted category result for the feature point, gt_i denotes the class of the true box, and N is the total number of samples.

$$L_{obj} = \frac{1}{M}\sum_j L_j = \frac{1}{N}\sum_j -[gt_j \cdot \lg(p_j) + (1-gt_j)\lg(1-p_j)] \tag{12}$$

where, p_j represents the predicted result of whether the feature point contains an object, gt_j denotes the class of the true box, and M is the total number of samples.

After obtaining the predicted bounding boxes for the feature points in the regression component, ES-IoU loss is used as the loss function. Section F provides a detailed explanation of the loss function calculation formula for the regression component.

$$L_{reg} = \frac{1}{Z}\sum_k L_k = \frac{1}{Z}\sum_k 1 - ESIoU^2 \tag{13}$$

The loss function of this paper is composed of these three parts combined into a multi-task loss function for joint optimization of training the network. Equation (14) represents the formula of the loss function in this paper.

$$L(\gamma) = \lambda L_{reg} + L_{obj} + L_{cls} \tag{14}$$

where, γ is the learning parameter of the network, and λ is the weight factor which is set to 5 according to reference [60].

F. ES-IOU LOSS FUNCTION

Some existing methods commonly utilize the CIoU loss function [30], defined as shown in Equations (15), (16), and (17).

$$CIoU = 1 - IoU + \frac{\rho^2(b, b_{gt})}{C^2} + av \tag{15}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w}{h}\right)^2 \tag{16}$$

$$a = \frac{v}{(1-IoU)+v} \tag{17}$$

where, IoU represents the intersection over union, b and b_{gt} respectively denote the centers of the predicted box and the ground truth box, ρ represents the Euclidean distance between these center points, C represents the diagonal length of the minimum rectangle that covers the predicted box and the ground truth box, w and w_{gt} respectively denote the widths of the predicted box and the ground truth box, h and h_{gt} respectively denote the heights of the predicted box and the ground truth box. Considering the gradient of penalty term v with respect to w and h, as shown in Equations (18) and (19).

$$\frac{\partial v}{\partial w} = -\frac{8}{\pi^2}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w}{h}\right) \times \frac{h}{h^2+w^2} \tag{18}$$

$$\frac{\partial v}{\partial w} = \frac{8}{\pi^2}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w}{h}\right) \times \frac{w}{h^2+w^2} \tag{19}$$

We can observe that:

In equation (16), v only reflects the difference in aspect ratio, rather than the actual relationship between w and w_{gt} or h and h_{gt}. That is, all instances with the property $\{(w = kw_{gt}, h = kh_{gt})\, k \in R^+\}$ have $v = 0$, which is inconsistent with reality;

In equation (18) and (19), we have $\frac{\partial v}{\partial w} = -\frac{h}{w}\frac{\partial v}{\partial h}$, where $\frac{\partial v}{\partial w}$ and $\frac{\partial v}{\partial h}$ have opposite signs. Therefore, at any given time, if one of these variables (w or h) increases, the other will decrease. This is unreasonable, especially when

$w < w_{gt}$ and $h < h_{gt}$ or $w > w_{gt}$ and $h > h_{gt}$;

Since v only reflects the difference in aspect ratio, the CIoU loss may optimize similarity in an unreasonable way,

hindering the effective reduction of the true differences between w, h, w_{gt}, and h_{gt}.

Addressing these two limitations, we propose a more efficient loss function called ES-IoU, defined as shown in Equations (20) and (21).

$$ESIoU = 1 - IoU + \frac{\rho^2\left(b, b_{gt}\right)}{C^2} + \frac{\rho^2\left(w, w_{gt}\right)}{C_w^2} + \frac{\rho^2\left(h, h_{gt}\right)}{C_h^2} + \gamma^2 \tag{20}$$

$$\gamma = \frac{\left(e^{-w_{gt}} - e^{-w}\right)^2 + \left(e^{-h_{gt}} - e^{-h}\right)^2}{2} + \varepsilon \tag{21}$$

where, C_w and C_h are the widths and heights of the minimum enclosing rectangle of the predicted and ground truth bounding boxes, ρ is the Euclidean distance between two points, ε is an offset added to prevent the situation where $\gamma = 0$, and the meanings of the other parameters remain consistent with CIoU. Additionally, each term is squared to accelerate the convergence speed of the loss function. Similarly, considering the gradient of γ with respect to w and h, as shown in Equations (22) and (23).

$$\frac{\partial \gamma}{\partial w} = \left(e^{-w_{gt}} - e^{-w}\right) \times e^{-w} \tag{22}$$

$$\frac{\partial \gamma}{\partial h} = \left(e^{-h_{gt}} - e^{-h}\right) \times e^{-h} \tag{23}$$

It can be observed that, in this case, the gradient function will not cause the problem of gradient explosion due to excessively small targets.

IV. EXPERIMENTS

A. DATASETS

This section conducted experiments on two widely used public datasets, CityPersons [12], CrowdHuman [34]. The results include ablation studies and performance comparisons with related methods.

1) CITYPERSONS

CityPersons [12] is a diversified pedestrian detection dataset evolved from the Cityscapes dataset. It contains a total of 5000 images with a size of 2048 × 1024 pixels, including 2975 images in the training set, 500 images in the validation set and 1525 images in the test set. In this paper, only the data from the "pedestrian" category, which represents walking, running, or standing human targets, is used for model training and testing. Additionally, as illustrated in Figure 6, the dataset is further divided based on different levels of occlusion: "Reasonable", "Bare", "Partial" and "Heavy".

2) CROWDHUMAN

The CrowdHuman [34], which is developed by MEGVII Technology, is specifically tailored for pedestrian detection. Most of the image data is obtained from Google searches. This extensive dataset includes 15,000 images in the training set, 5,000 images in the test set, and 4,370 images in the validation set. In total, the training and validation sets comprise 470,000 instances, with an average of around 23 people per image. Diverse occlusion scenarios are captured simultaneously. Each human instance is meticulously annotated with bounding boxes for the head, visible area, and the entire body.

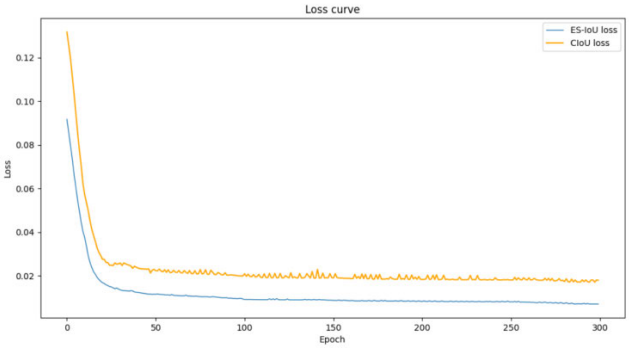

FIGURE 5. The loss curve during training on the CityPersons dataset.

FIGURE 6. Samples for four occlusion levels.

B. EXPERIMENTAL ENVIRONMENT AND IMPLEMENTATION DETAILS

The experiment was conducted using a Parsai server with Ubuntu 20.04 operating system, Intel Core Xeon Platinum 8373 processor (36 cores, 2.6 GHz), 64GB RAM, and NVIDIA GeForce RTX 4090 GPU (24GB VRAM). The training process was implemented using open-source Python language and Pytorch.

During the training phase, images from the CrowdHuman datasets [34] were resized to 640 × 640 pixels, with each training batch containing 8 images. The training iterations were set to 150. Input images were randomly preprocessed using operations such as color distortion, image flipping, and image scaling. The Adam optimizer [35] was used with an initial learning rate of 0.01, which was adjusted to 0.0001 from the 30th epoch onwards.

During the testing phase, images from the CrowdHuman [35] and CityPersons [12] datasets were resized to 1280 × 1280 and 2048 × 1024 pixels, respectively, without any additional preprocessing operations.

C. EVALUATION METRICS

MR (Miss Rate) is an evaluation metric used to describe the results of human body detection, where lower values indicate better performance. The accuracy of human body detection is primarily reflected in two aspects: detecting as many human targets as possible while minimizing false positives. Therefore, in the evaluation of pedestrian detection performance, MR and FPPI (False Positives per Image) are usually considered together, and it is often necessary to adjust the decision threshold (or confidence score) in the detection algorithm to achieve a balance between the two. If a lower decision threshold is set, MR decreases and FPPI increases; conversely, if a higher decision threshold is set, MR increases and FPPI decreases. Hence, to better represent the performance of human body detection under different decision threshold conditions, Dollar [45] proposed calculating the logarithmic average of the MR within a certain range of FPPI as a quantitative metric, referred to as the log-average miss rate (LAMR). The calculation formula is as follows:

$$LAMR = exp(\frac{\sum_{i=1}^{N}(MR(FPPI_i)}{N}) \tag{24}$$

where $FPPI_i$ represents the FPPI value corresponding to the selected sampling point i, and N denotes the number of sampling points. To better reflect the miss rate of the detector under low false positive conditions and to facilitate fair comparisons with existing methods, FPPI is sampled at intervals of $10^{0.25}$ in the range of $[10^{-2}, 10^{0}]$. The logarithmic average miss rate (LAMR) in this state is referred to as the miss rate and denoted as MR^{-2} in this paper.

AP (average precision) represents the average accuracy of all image detections belonging to a certain class. The calculation formula is as follows:

$$AP = \frac{\sum_n Precision}{n} \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \tag{26}$$

where n represents the total number of images belonging to a certain class. A higher AP value indicates better performance of the detection model.

This paper follows the division criteria for objects of different scales based on the COCO dataset [44], as shown in Table 1. Pedestrians with an area less than or equal to 32×32 pixels are categorized as small scale; pedestrians with an area greater than 32×32 pixels and less than 96×96 pixels are categorized as middle scale and pedestrians with an area greater than or equal to 96×96 pixels are categorized as large scale. For the CityPersons dataset, the official evaluation standard of Miss Rate (MR^{-2}) is used for assessment, where a lower value indicates better detection performance. Similarly, to illustrate the detection performance of the proposed model, this article discusses MR values for five different scenarios on CrowdHuman, with criteria primarily based on varying degrees of occlusion and scale. The division criteria for different degrees of occlusion based on the CityPersons dataset are shown in Table 2.

TABLE 1. The scaling criteria for objects in the COCO dataset.

Region	Object Scale
area <32×32 pixels	Small
32×32 <area <96×96 pixels	Middle
96×96 pixels <area	Large

TABLE 2. The division criteria for certain subsets within the CityPersons dataset.

Subset	Pedestrian Height	Occlusion Level
Bare	>50 PXs	0.1≤ occlusion
Reasonable	>50 PXs	occlusion <0.35
Partial	>50 PXs	0.1< occlusion ≤0.35
Heavy	>50 PXs	0.35< occlusion ≤0.8

TABLE 3. Module validation ablation experiment results on module validation (MR^{-2}%).

Method	Reasonable	Heavy	Small
Baseline	13.9	49.2	16.9
Baseline+TGCM	13.1	48.1	13.5
Baseline+TGCM+Self-Attention	11.2	46.6	15.3
Baseline+TGCM+Channel Attentiom	11.5	46.9	13.7
Ours	**10.1**	**45.3**	**11.7**

D. ABLATION EXPERIMENTS

To verify the effectiveness of the feature enhancement module, experiments were conducted by comparing detectors that exclude the feature enhancement module and global contextual information as the baseline. These experiments were carried out on the CityPersons dataset, and the evaluation metrics used were MR^{-2} values under three scenarios: Reasonable, Heavy, and Small. The experimental results are shown in Table 3. From the module verification experiment results in Table 3, the following observations can be made: Firstly, the feature fusion strategy retains most of the features of medium and small-scale pedestrians, but the introduction of a lot of background noise at the same time prevents the overall detection performance from being optimized. Secondly, the self-attention module, built upon the feature fusion strategy, enhances the features of small-scale pedestrians by utilizing contextual information of the features while suppressing background information. Thirdly, the channel attention module, due to insufficient feature information, does not yield ideal overall detection performance but significantly improves the detection performance of small-scale pedestrians. This demonstrates the effectiveness of the channel attention module in enhancing the features of small-scale pedestrians through nonlinear modeling of channel correlations. Finally, with all three sub-modules working together, the proposed model not only improves the overall detection accuracy but also optimizes the detection of medium and small-scale pedestrians.

In summary, compared to the baseline, the proposed model achieved a 2.8% improvement on Reasonable,

TABLE 4. Comparison experiments between different channel attentions and self-attentions(MR^{-2}%).

Method	Reasonable	Heavy	Small
Baseline+TGCM	13.1	48.1	13.5
Baseline+TGCM+CA(SENet)+SA	11.5	46.3	12.6
Baseline+TGCM+CA(CBAM)+SA	11.9	46.9	13.1
Baseline+TGCM+CA(ECA-Net)+SA	11.7	47.0	12.9
Baseline+TGCM+SA+CA(CBAM)	10.8	45.7	12.3
Baseline+TGCM+SA+CA(ECA-Net)	10.6	45.9	12.1
Baseline+TGCM+SA+CA(SENet)**(Ours)**	**10.1**	**45.3**	**11.7**

demonstrating the enhancement effect of the proposed modules on pedestrian detection. Meanwhile, the detection accuracy for small-scale pedestrians improved by 5.2%, validating the effectiveness of the designed modules for small-scale pedestrian detection.

To validate the impact of different types of channel attention and the order of channel attention and self-attention on model detection accuracy, we conducted comparative experiments, the results are shown in Table 4, where CA stands for channel attention and SA stands for self-attention. The experimental results indicate that employing TGCM and then self-attention followed by the channel attention proposed in SENet yields the best results. Although in the CBAM [36] paper, the channel attention proposed in CBAM (which includes MaxPool and AvgPool) performs better than the channel attention proposed in the SENet (Squeeze-and-Excitation) method, however, CBAM only conducted classification experiments on ImageNet-1K dataset and did not verify its effectiveness in pedestrian detection. For pedestrian detection tasks, our proposed approach of using TGCM and self-attention followed by the channel attention from SENet achieves the best results. The main reason may be that the channel attention proposed in CBAM includes an additional MaxPool operation compared to the channel attention proposed in SENet [33]. Max Pooling can lead to inconsistent detection performance for pedestrians of different scales, especially for smaller pedestrians, which may decrease detection accuracy due to scale variations. Similarly, in the small-scale pedestrian detection task of this paper, using the channel attention proposed in SENet yields better results than the channel attention proposed in ECA-Net [37].

To verify the proposed ES-IoU loss function, the performance of several mainstream loss functions in small object detection was compared. The experimental results are shown in Table 5. From the table, it can be observed that the ES-IoU proposed in this paper performs well in most metrics. It only slightly lags behind EIoU in the Reasonable metric, mainly because EIoU minimizes the differences in width and height between predicted and ground truth boxes. In contrast, the designed loss function ES-IoU considers the characteristics of small-scale pedestrians comprehensively and plays a role in accelerating convergence by alleviating gradient explosion. This is evident in the excellent performance on the Heavy and Small metrics, indicating that ES-IoU focuses more on high-quality predicted boxes. Overall, considering all metrics, ES-IoU is more suitable for small object detection.

TABLE 5. Loss function comparison experiment results (MR^{-2}%).

Method	Reasonable	Heavy	Small
DIoU[30]	11.3	46.8	13.9
CIoU[30]	10.9	46.3	13.1
EIoU[46]	**9.9**	45.8	12.5
ES-IoU (Ours)	10.1	**45.3**	**11.7**

TABLE 6. Transformer global context ablation experiments.

Method	Reasonable	Heavy	Small
F_2	12.3	47.2	13.2
F_3	12.1	47.1	13.0
F_4	10.8	45.9	12.6
F_5	10.6	45.7	12.4
F_6	11.1	46.5	13.2
F_2+F_3	11.2	46.3	12.8
F_4+F_5	**10.1**	**45.3**	**11.7**

To further validate the impact of the Top-Down Global Context Module (TGCM) on small object detection, we conducted comparative experiments by adding this module at different positions in the backbone network. The experimental results are shown in Table 6.

From the experimental results, it can be observed that the Transformer-based Global Context Information Module (TGCM) exhibits significant improvements when applied in deeper layers, while its performance in shallower layers is not particularly ideal. Interestingly, as the network depth increases, the performance of the network tends to degrade. As shown in the table, when TGCM is added only on a single layer, the best results for Reasonable and Small metrics are achieved when added at the F_5 layer. The inferior performance of the F_6 layer compared to the F_5 layer is mainly due to that as the feature pyramid goes higher, i.e., deeper layers, the scale of the feature maps becomes smaller, resulting in fewer features containing small object regions and thus deteriorating small object detection performance.

Furthermore, combining experiments and analysis, we attribute the inferior performance of this module in F_2, F_3, and F_4 compared to F_5 to the following reasons. Firstly, the quality of features learned in shallower layers is not high. The features learned in shallower layers mostly consist of easily learned low-level features such as texture and appearance of the target, and some channel branches even learn features like background noise. In contrast, deeper layer features tend to focus more on human body targets. This module enables the interaction of feature information learned by each channel with the global context to make better use of global contextual information. Secondly, there are not enough channels in shallower layers. The shallow layers have too few feature channels, resulting in insufficient feature information for interaction. Even if some feature information about small objects is learned in shallow layers, the TGCM module finds it challenging to interact with the global context and learn more discriminative features.

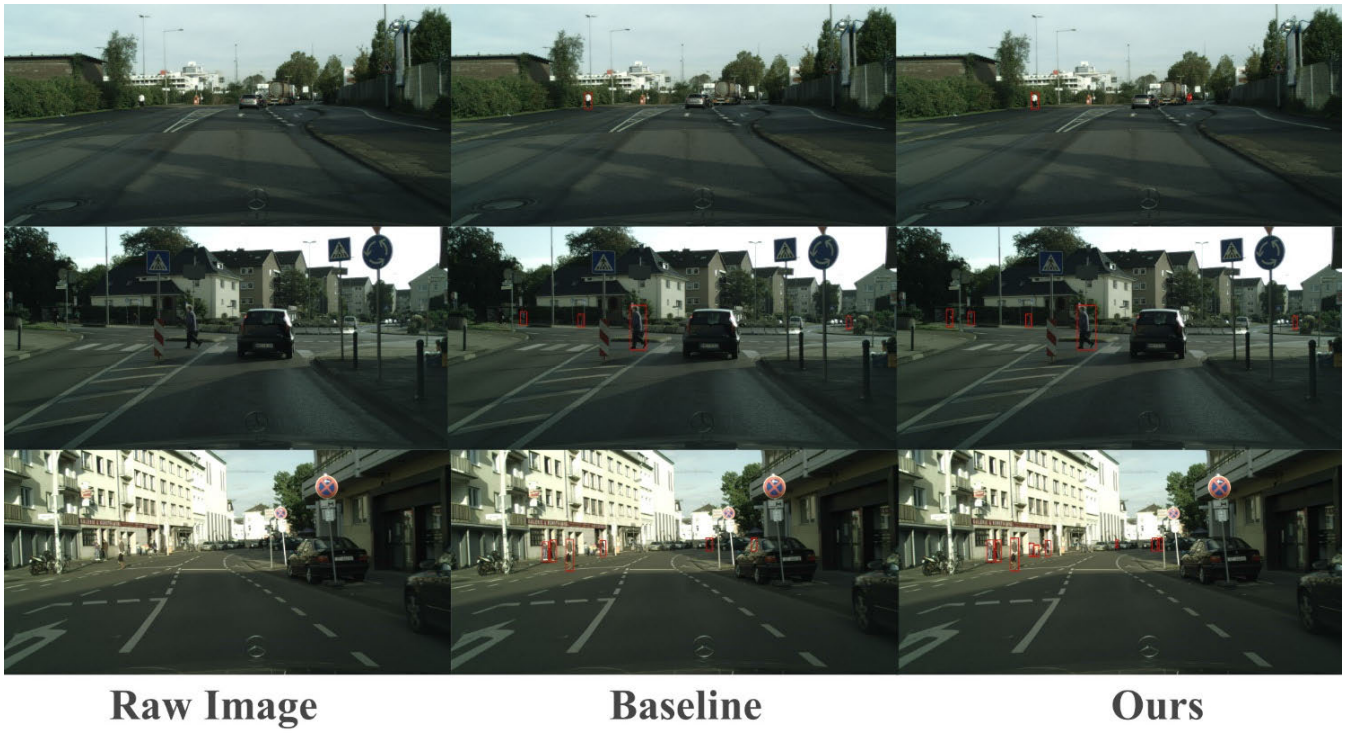

FIGURE 7. Illustration of the results for different methods.

Additionally, it is observed that adding TGCM to both F_4 and F_5 layers simultaneously and fusing information leads to better experimental results. This is because it effectively integrates human body target features. Therefore, this paper chooses to add TGCM to both F_4 and F_5 layers simultaneously.

E. COMPARISON WITH STATE-OF-THE-ART METHODS

To assess the performance of the algorithm proposed in this paper, we conducted comparative experiments on different subsets of the CityPersons dataset, utilizing MR^{-2} as the evaluation metric. Thirteen typical methods from the CityPersons dataset were chosen for comparison, and the results of these experiments are presented in Table 7.

From Table 7, we can conclude that the model proposed in this paper demonstrates optimal performance on small-scale pedestrians. Compared to the currently leading algorithm, it shows an improvement of 1.7%. Moreover, our proposed model also outperforms others on the Heavy dataset and demonstrates commendable performance across subsets such as Bare, Reasonable, and Partial. These experimental results strongly support the effectiveness of the model introduced in this study.

F. GENERALIZATION EXPERIMENT

To assess the generalization capabilities of our proposed model, we conducted experiments on the CrowdHuman dataset. Evaluation metrics such as Average Precision (AP), Recall, Miss Rate (MR^{-2}), and Small AP values were employed.

TABLE 7. Comparison of miss rate with existing methods on CityPersons (MR^{-2}%).

Method	Bare	Reasonable	Partial	Heavy	Small
RepLoss[48]	7.6	13.2	16.8	56.9	42.6
TLL[13]	10.0	15.5	17.2	53.6	-
OR-CNN[49]	6.7	12.8	15.3	55.7	42.3
CSP[15]	7.3	11.0	10.4	49.3	16.0
PAPER[50]	7.9	10.6	10.2	50.2	14.3
NOH-NMS[51]	6.6	10.8	11.2	53.0	-
AP^2M[41]	6.2	10.4	9.7	48.6	15.3
SML[42]	-	12.3	-	-	19.3
MAPD[38]	**6.1**	**9.7**	9.9	46.4	-
Beta RCNN[52]	6.4	10.6	10.3	47.1	-
SMPD[39]	6.5	9.9	**9.0**	45.6	-
DHRNet[40]	-	10.4	-	-	13.4
NMS-Ped[43]	-	10.1	-	-	-
Ours	6.3	10.1	9.3	**45.3**	**11.7**

The results are summarized in Table 8. Notably, our model achieved a detection accuracy of 22.3% for small-scale pedestrians on CrowdHuman dataset, showcasing a notable improvement of 5.8% over baseline algorithms. Furthermore, the model exhibited optimal performance in terms of MR^{-2} and AP. These experimental results confirm the robust generalization performance of our proposed model.

G. VISUALIZATION

We compared the visual results of our improved model with the baseline network, as shown in Figure 7, it is evident that the improved model exhibits significant improvements in detecting small-scale pedestrians compared to the original

TABLE 8. Comparison of miss rate with existing methods on CrowdHuman.

Method	AP	Recall	MR^{-2}	Small AP
FPN[17]	83.1	90.6	52.4	-
Yolov4[47]	75.3	92.5	64.9	16.4
FPN+Soft-NMS[53]	83.9	91.7	52.0	-
FPN+ FRCNN[17]	84.5	90.2	50.4	-
FPN+AdaptiveNMS[54]	84.7	91.3	49.7	-
RelationNet[26]	81.6	-	48.2	-
RFB-Net[55]	78.3	94.1	65.2	-
RetinaNet[56]	80.8	93.8	63.3	16.5
FCOS+AEVB[58]	-	-	47.7	-
YOLO-CS[57]	81.9	**95.3**	41.9	-
Ours	**89.1**	94.1	**41.6**	**22.3**

model, with noticeable reductions in missed detections and false alarms, demonstrating superior performance.

V. CONCLUSION

In this work, we propose a small-scale pedestrian detection algorithm based on multi-scale structural perception and global contextual information. Firstly, to integrate feature information of small objects at different scales and quickly locate pedestrians, a multi-scale structural perception module is proposed. The feasibility of this module is validated through experimental results and visual analysis, showing that it enhances the network's focus on small-scale pedestrian features. Secondly, to better utilize contextual information, the paper leverages the advantages of capturing long-distance dependencies using Transformer structures and proposes a global contextual information module. This module enables interaction and learning among different channels, considering that small-scale pedestrian features are prone to confusion with background information. Through the joint modeling of spatial and channel correlations of feature maps using self-attention and channel attention modules, it enhances small-scale pedestrian features while suppressing background information. Lastly, a loss function more suitable for small-scale pedestrian detection, namely ES-IoU loss function, is proposed, which effectively accelerates the convergence speed of the model. Extensive experimental results demonstrate significant improvements in metrics such as recall and average precision, validating that the proposed method is more suitable for small-scale pedestrian detection in complex scenarios.

REFERENCES

[1] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[2] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4073–4082.

[3] 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)*, Jun. 2015, pp. 5079–5087.

[4] 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)*, Jun. 2016, pp. 1259–1267.

[5] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3506–3515.

[6] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3258–3265.

[7] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli, "Pushing the limits of deep CNNs for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1358–1368, Jun. 2018.

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

[9] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[10] 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.

[11] 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)*, 2016, pp. 443–457.

[12] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4457–4465.

[13] 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)*, 2018, pp. 536–551.

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

[15] 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)*, Jun. 2019, pp. 5182–5191.

[16] 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.*, Oct. 2016, pp. 21–37.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[18] C. Lu, S. Wu, C. Jiang, and J. Hu, "Weak harmonic signal detection method in chaotic interference based on extended Kalman filter," *Digit. Commun. Netw.*, vol. 5, no. 1, pp. 51–55, Feb. 2019.

[19] X. Luo, J. Li, W. Wang, Y. Gao, and W. Zhao, "Towards improving detection performance for malware with a correntropy-based deep learning method," *Digit. Commun. Netw.*, vol. 7, no. 4, pp. 570–579, Nov. 2021.

[20] 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. 2018.

[21] 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)*, 2016, pp. 354–370.

[22] J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3872–3884, 2021.

[23] X. Xie and Z. Wang, "Multi-scale semantic segmentation enriched features for pedestrian detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2196–2201.

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

[25] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1951–1959.

[26] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[27] Y. Tan, H. Yao, H. Li, X. Lu, and H. Xie, "PRF-Ped: Multi-scale pedestrian detector with prior-based receptive field," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 6059–6064.

[28] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, Oct. 2016, pp. 516–520.

[29] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
[30] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, no. 7, pp. 12993–13000.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and A. N. Gomez, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.
[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[34] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[38] Y. Wang, C. Han, G. Yao, and W. Zhou, "MAPD: An improved multi-attribute pedestrian detection in a crowd," *Neurocomputing*, vol. 432, pp. 101–110, Apr. 2021.
[39] H. Jiang, S. Liao, J. Li, V. Prinet, and S. Xiang, "Urban scene based semantical modulation for pedestrian detection," *Neurocomputing*, vol. 474, pp. 1–12, Feb. 2022.
[40] M. Ding, S. Zhang, and J. Yang, "Learning a dynamic high-resolution network for multi-scale pedestrian detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9076–9082.
[41] M. Liu, C. Zhu, J. Wang, and X.-C. Yin, "Adaptive pattern-parameter matching for robust pedestrian detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2154–2162.
[42] J. Wu, C. Zhou, Q. Zhang, M. Yang, and J. Yuan, "Self-mimic learning for small-scale pedestrian detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Seattle, WA, USA, Oct. 2020, pp. 2012–2020.
[43] Z. Luo, Z. Fang, S. Zheng, Y. Wang, and Y. Fu, "NMS-loss: Learning with non-maximum suppression for crowded pedestrian detection," in *Proc. Int. Conf. Multimedia Retr.*, Taipei, Taiwan, Aug. 2021, pp. 481–485.
[44] T. Lin, M. Maire, and S. J. Belongie, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[45] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[46] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, Sep. 2022.
[47] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[48] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.
[49] 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)*, 2018, pp. 637–653.
[50] Y. Chen, W. Xie, H. Liu, B. Wang, and M. Huang, "Multi-feature fusion pedestrian detection combining head and overall information," *J. Electron. Inf. Technol.*, vol. 44, no. 4, pp. 1453–1460, 2022.
[51] 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*, Oct. 2020, pp. 1967–1975.
[52] Z. Xu, B. Li, Y. Yuan, and A. Dang, "Beta R-CNN: Looking into pedestrian detection from another perspective," 2022, *arXiv:2210.12758*.
[53] 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. (ICCV)*, Oct. 2017, pp. 5562–5570.
[54] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6452–6461.
[55] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 404–419.
[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[57] H.-H. Xu, X.-Q. Wang, D. Wang, B.-G. Duan, and T. Rui, "Object detection in crowded scenes via joint prediction," *Defence Technol.*, vol. 21, pp. 103–115, Mar. 2023.
[58] Y. Zhang, H. He, J. Li, Y. Li, J. See, and W. Lin, "Variational pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11617–11626.
[59] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
[60] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.0843*.

HAO GAO is currently pursuing the bachelor's degree majoring in computer science and technology with Jiangsu University of Science and Technology, Zhenjiang, China. He is interested in algorithms related to object detection.

SHUCHENG HUANG received the bachelor's and master's degrees in computer application from China University of Mining and Technology, Xuzhou, China, in 1991 and 2001, respectively, and the Ph.D. degree in computer application from Southeast University, Nanjing, China, in 2007. He is currently a Professor with Jiangsu University of Science and Technology, Zhenjiang, China. His current research interests include computer vision and multimedia analysis.

MINGXING LI was born in Jiangsu, China, in 1989. He received the bachelor's and master's degrees in computer science and technology from Jiangsu University of Science and Technology, Zhenjiang, China, in 2011 and 2014, respectively. His current research interests include computer vision and multimedia analysis.

TIAN LI was born in Gansu, China, in 1988. She received the bachelor's and master's degrees in computer science and technology from Jiangsu University of Science and Technology, Zhenjiang, China, in 2010 and 2013, respectively. Her research interests include deep learning and computer vision.