Received 26 January 2023, accepted 9 February 2023, date of publication 14 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244981

RESEARCH ARTICLE

PF_YOLOv4: An Improved Small Object Pedestrian Detection Algorithm

KAIHUI LI[1,2], YUAN ZHUANG[1,2], JINLING LAI[1], AND YUNHUI ZENG[1,2]
[1]Faculty of Computer Science and Technology, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China
[2]Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China

Corresponding author: Yunhui Zeng (zengyh@sdas.org)

This work was supported by the Shandong Provincial Key Research and Development Program under Grant 2019JMRH0109.

ABSTRACT With the development of deep convolutional neural networks, the effect of pedestrian detection has been rapidly improved. However, there are still many problems in small target pedestrian detection, for example noise (such as light) interference, target occlusion, and low detection accuracy. In order to solve the above problems, based on YOLOv4 algorithm, this paper proposes an improved small target pedestrian detection algorithm named PF_YOLOv4. The algorithm is improved in three aspects on the basis of the YOLOv4 algorithm: firstly, a soft thresholding module is added to the residual structure of the backbone network to perform noise reduction process on interference factors, such as light to enhance the robustness of the algorithm; secondly, the depthwise separable convolution replaces the traditional convolution in the YOLOv4 residual structure, to reduce the number of network model parameters; finally, the Convolutional Block Attention Module (CBAM) is added after the output feature map of the backbone network to enhance of the network feature expression. Experimental results show that the PF_YOLOv4 algorithm outperforms most of the state-of-the-art algorithms in detecting small target pedestrians. The mean Average Precision (mAP) of the PF_YOLOv4 algorithm is 2.35% higher than that of the YOLOv4 algorithm and 9.67% higher than that of the YOLOv3 algorithm, while the detection speed is slightly higher than that of YOLOv4 algorithm.

INDEX TERMS Small target pedestrian detection, soft thresholding, depthwise separable convolution, convolutional block attention module.

I. INRTODUCTION

Small target pedestrian detection is a basic task in pedestrian detection, which mainly studies the accurate identification and localization of small target pedestrians from images or sequential images. It is not only employed in various practical scenarios for example, intelligent transportation systems and security monitoring, but also provides a theoretical basis for research hotspots such as human behavior recognition and motion understanding. Recent years have witnessed a spurt of progress in deep learning theories and the continuous enhancement of computing hardware, deep neural networks have achieved the desired effect in vision tasks, and also been employed in pedestrian detection. Small object pedestrian detection has gradually become one of the research hotspots in computer vision [1], [2], [3]. Small target pedestrians refer to pedestrians with a relatively small proportion in the input videos or image data. The MS COCO [4] dataset defines targets with pixels less than 32 × 32 as small targets. Considering the special proportion of pedestrians, this paper considers pedestrians whose heights are less than 32 pixels are treated as small target pedestrians.

Although pedestrian detection achieves impressive results on large or medium-sized pedestrians, the performance on small target pedestrians is a long way from satisfactory [5], [6]. There are many challenges for example noise (such as light) interference, small target pedestrians are occluded, and small target pedestrians are too small, so that small target pedestrians lack sufficiently detailed appearance information, and their feature expression ability is relatively low. In the

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

feature extraction process, only a small number of features can be extracted, resulting in low detection accuracy of small target pedestrians.

To solve these issues, this paper designed an novel small target pedestrian detection algorithm named PF_YOLOv4, to improve the detection effect on the small target pedestrian detection task. Firstly, a soft thresholding module is added to the residual structure of the backbone network. Through this module in parallel with the residual structure, it can judge whether there is noise interference in pedestrian detection of small targets, and perform noise reduction processing on interference factors such as light, so as to improve the detection accuracy. After the soft thresholding module is added, the algorithm becomes more complicated and the detection speed becomes slower. Secondly, in order to improve the running situation, the paper uses depthwise separable convolution, which reduces the parameters and computation load of the detection algorithm. Finally, CBAM is added after the output feature map of the backbone network. Through the channel attention module and the spatial attention module in the CBAM, the algorithm's ability to extract global features and local features is improved. In the scene where the small target pedestrian is occluded and the small target pedestrian is small, the detection effect has been greatly improved, and the occluded small target pedestrian and particularly small pedestrian can be better identified.

This paper has the following three contributions:

1) A soft thresholding function is introduced to deal with noise such as light in the image. So small target pedestrians can be more accurately identified.
2) The depthwise separable convolution is introduced in the backbone network. It is able to decrease the amount of network parameters and improve the network training speed.
3) CBAM is added after the output feature map of the backbone network. The extraction of local and global features of small target pedestrians and the detection accuracy of small target pedestrians is enhanced.

Experiments of the proposed PF_YOLOv4 on MS COCO and other public datasets are made. The result demonstrate that the PF_YOLOv4 algorithm is better than other algorithms in detecting small target pedestrians.

In Section II, related work and literature on pedestrian detection and small object pedestrian detection are discussed. In Section III, the structure of YOLOv4 is introduced. In Section IV, the specific method PF_YOLOv4 of this paper is proposed and fully described. In Sections V, extensive experiments are performed, and the method is put in to databases about pedestrian, which is used to test feasibility and performance. Finally, Section VI concludes this paper.

II. RELATED WORK

With the introduction of deep convolutional neural networks (CNNs), more CNN-based deep learning models have enhanced the performance of object pedestrian. Pedestrian detection methods based on CNN are mainly classified into two types: two-stage algorithm [7] and one-stage algorithm [8].

Two-stage algorithm, which first generate regional recommendations and then perform classification and regression [9]. The typical representation of such algorithms is the R-CNN series. R-CNN [10] is the first algorithm to successfully apply deep learning to target detection, but it is too slow in the reality problem. He et al. [11] proposed SPPNet, adding SPP layer between the final convolution layer and full connection layer of R-CNN, which can input arbitrary pictures and improve the detection speed. Girshick [12] proposed Fast-RCNN on the basis of R-CNN, further performed ROIPooling operation on each candidate region, then used softmax to perform multi-classification operation, and finally used regression model to adjust the size to determine the boundary position. Ren et al. [13] proposed Faster R-CNN, which integrates extracted feature, bounding box regression, candidate regions and classification into one network, which promotes the improvement of comprehensive performance. Lai et al. [14] proposed the MSRCR-IF algorithm, which improved the detection accuracy of pedestrians in low lightness by adjusting the Region Proposal Network (RPN) and removing the instance mask branch. The region nomination detection algorithm has the advantage of high detection accuracy, but it requires more time cost, it is difficult to realize real-time detection.

One-stage algorithm, ignoring the process which produces the chosen area in candidate framework, and it also produces the category probability and location coordinate object value which need to detect, identify, and classify [9]. The representatives of such algorithms are mainly SSD and YOLO series. Liu et al. [15] proposed the SSD, which mainly uses the feature layer of the pyramid structure to extract the feature image, and then uses the softmax and position regression to classify the feature map extracted by the multi-convolution layer. However, SSD uses low-level feature information, which makes the information of different scales blend poorly. Fu et al. [16] also proposed the DSSD, which uses Resnet101 [17] as the basic network to increase image feature fusion, so as to enhance the ability to extract feature information. Jeong et al. [18] proposes R-SSD, which enhances the connection of feature information between different convolutional layers and improves the detection effect of small targets. Cao et al. [19] proposed that Feature-fused SSD integrates the feature information of high-level and low-level convolutional neural networks at the same time, which improves the detection effect of small targets. For fast feature fusion of different convolutional layers, Li and Zhou [20] proposed FSSD. Different from the complex network of FPN, FSSD directly converts the operation to bilinear interpolation of different convolutional layers into the same scale, and then fuses different feature information, and obtains a good detection effect.

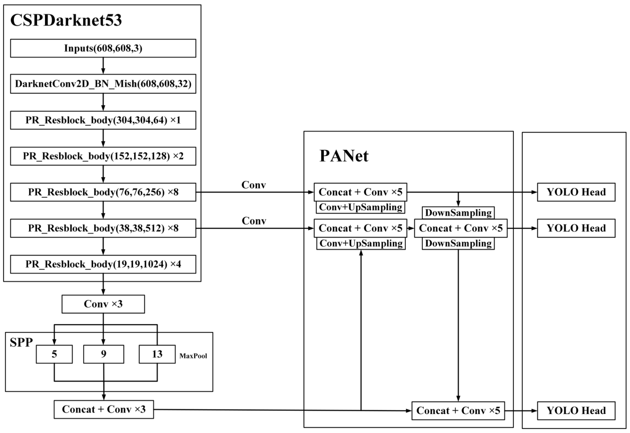


FIGURE 1. The structure diagram of YOLOv4.

Redmon et al. proposed YOLO [21], which can predict multiple bounding box (BBox) positions and categories at one time. In 2016, J. Redmon and Farhadi [22] proposed YOLOv2, which utilizes K-means clustering to compute better anchor templates in the training set. However, YOLOv2 uses the features of the last convolutional layer, which loses a lot of information. In 2018, Redmon made some improvements on YOLOv2 and proposed YOLOv3 [23].The darknet-53 network structure is used to replace the original darknet-19, and the feature pyramid network structure is used to realize multi-scale detection. The classification method uses logistic regression instead of softmax, which ensures the accuracy of target detection while taking into account the real-time performance. But, the performance of YOLOv3 did not effectively combined with BBox. Bochkovskiy et al. [24] used CSPDarknet53 [25] as the backbone network in YOLOv4 to get a better performance of detection accuracy and speed, and added SPP blocks to improve the size of the receptive field. In 2021, Boyuan and Muqing [26] proposed combining the Spatial Pyramid Pooling (SPP) network and the K-means clustering algorithm with the YOLOv4 model, and used the Mish activation function in the neck of the model to better solve the impact of the occlusion problem on target detection. In 2021, to relieve the impact of illumination on pedestrian detection, Cao et al. [27] designed a new multispectral channel feature fusion (MCFF) module based on the YOLOv4 algorithm to integrate color and heat flow information under different lighting conditions, so it improved pedestrian detection accuracy.

III. PRELIMINARY WORK

Owing to the deep network structure, multi-scale feature fusion and anchor frame mechanism, YOLOv4 has a faster detection speed while maintaining a higher accuracy. The performance of YOLOv4 on MS COCO and other datasets exceeds that of SSD, RetinaNet and other algorithms. Although the detection speed of YOLOv5 is faster than that of YOLOv4, the overall performance of YOLOv4 is better. So, this paper chooses YOLOv4 as the basic algorithm.

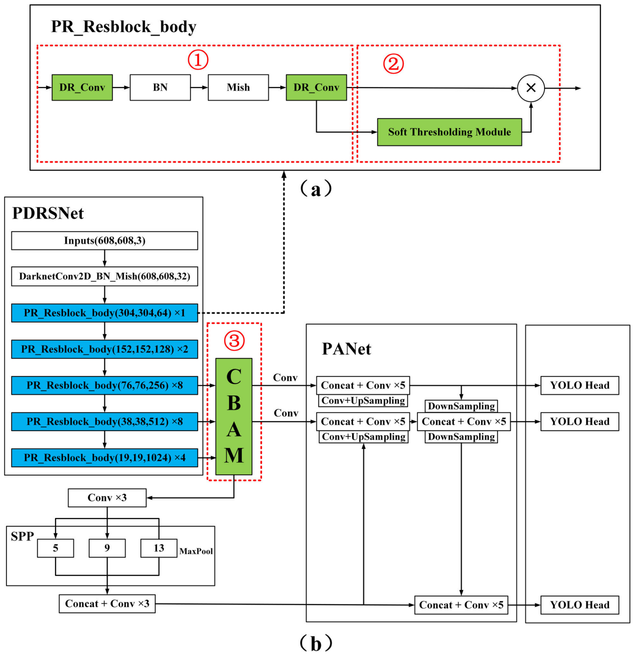


FIGURE 2. An overview of PF_YOLOv4. (a)PR_Resblock_body module, (b)The structure diagram of PF_YOLOv4. In ①, DR_Conv stands for depthwise separable convolution module, ② is soft thresholding module, ③ is CBAM. There is a total of 23 PR_Resblock_body in the backbone network. The traditional convolution in PR_Resblock_body is replaced by a depthwise separable convolution, and a soft thresholding module is added. PF_YOLOv4 backbone network is replaced from CSPDarknet53 to parallel deep contraction residual network (PDRSNet).

The YOLOv4 structure is shown in Figure 1. The YOLOv4 algorithm consists of four parts: the input layer, the Backbone layer, the Neck layer and the PANet layer. The input layer is the input of the original image, and the image data is preprocessed, which is mainly used to increase the image features of small objects [24]. The Backbone layer uses CSPDarknet53, which contains 1 CBM layer and 5 CSP modules. The Neck layer is a feature enhancement module, which is mainly composed of CBL components, SPP modules [11] and FPN + Path Aggregation Network (PAN) [28] methods. The PANet layer is a multi-scale feature fusion module. And it fuses multi-level feature images up and down to fully display the feature information of each position of the image.

The YOLOv4 algorithm adopts multi-scale prediction, and divides the input image into $S \times S$ small grids on average. Each grid uses 3 anchors, and each anchor information is represented by $T(x, y, w, h, c)$. Among them, (x, y) represents the center coordinates, (w, h) represents the width and height, and c represents the confidence. The output dimension of each anchor is $S \times S \times 3 \times (4 + 1 + C_1)$, and C_1 represents the category. This paper only takes the small target pedestrian as the target detection object, so C_1 is set to 1.

IV. METHODS OF THIS PAPER

Figure 2 illustrates the PR_Resblock_body module and PF_YOLOv4 structure. Three modules are added on the basis of YOLOv4: soft thresholding module, depthwise separable

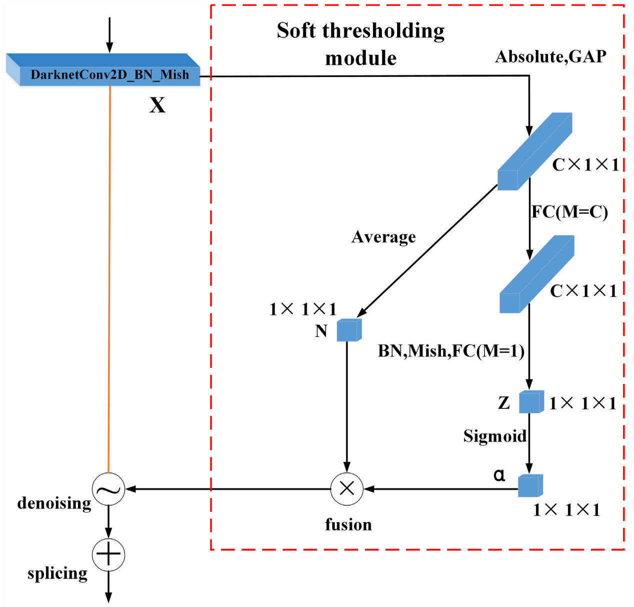


FIGURE 3. The structure diagram of PRNet. BN (Batch Normalization), GAP (Global Average Pooling), FC (Fully Connected Layer). Among them, M represents the number of nerve cells in the fully connected layer. M = C means that the number of nerve cells is equal to the number of channels in the previous feature map, and M = 1 means that the number of nerve cells is 1. Absolute means to take the absolute value of each element in the feature map, and Average means to find the average value of all elements in the feature map.

convolution module and CBAM. First, a soft thresholding module is added to process noise, then a depthwise separable convolution module is added to increase the algorithm detection speed, and finally a CBAM is added to enhance the feature extraction capability of the network. These three modules are described in detail below.

A. SOFT THRESHOLDING MODULE

The interference of non-pedestrian objects is often encountered in the process of small target pedestrian detection, especially the light interference, which leads to low detection accuracy of small target pedestrians. These factors which are affecting small target pedestrian detection can be attributed to noise. As a general target detection algorithm, YOLOv4 does not have a module for noise reduction. From the perspective of image noise reduction to increase the accuracy, it is necessary to add noise reduction method in the process of small target pedestrian detection. Soft threshold function is a commonly used concept of signal noise reduction, and it reduces the value of the signal segment to “ 0 ”. In the deep residual shrinking network, the soft threshold is a branch parallelized, and noise reduction is performed through this branch. Whether to reduce noise or how to reduce noise needs to be determined after the calculation through the soft thresholding formula. This paper draws on the idea of soft thresholding in deep residual shrinkage networks [29], and proposes a parallel residual network combined with soft thresholding, named PRNet. When extracting the feature information of small target pedestrians, noise such as light is removed.

The PRNet structure is shown in Figure 3. In the soft thresholding module, first, the absolute values of all features in the input feature map are obtained. Then, after global pooling and average pooling, a feature map is obtained, denoted as M. In the left path, the mean N of the absolute values of the feature M is obtained and fed into a small and fully connected network. In the right path, the fully connected network uses the sigmoid function as the last layer. The output result is scaled to the range of (0,1), which is represented by α. Finally, the threshold is $\alpha \times N$, denoted as x. Input the soft threshold x into the soft threshold formula to determine whether to perform noise reduction. The soft threshold formula is shown in (1):

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \leq x \leq \tau \\ x + \tau & x < -\tau \end{cases} \tag{1}$$

The x means the input and y means the output, τ is the threshold. The threshold is a positive and small number. If the threshold is greater than the absolute value of all input attributes, output y can only be zero. In this case, the soft threshold is meaningless, indicating that the noise has little effect on the detection accuracy and can be ignored. The result of the residual network and the denoised feature map is spliced together, and the factors affecting the small target pedestrian are eliminated through noise reduction processing. In the process of extracting the feature information of small target pedestrians, the noise interference such as light in the feature map is removed through the parallel soft thresholding module.

B. DEPTHWISE SEPARABLE CONVOLUTION MODULE

Traditional convolution is used in the residual structure of YOLOv4. It convolves and multiples the input of each channel with the corresponding convolution kernel, then accumulates the results, and outputs the feature map. After the soft thresholding module is added, the algorithm becomes more complex, and the detection speed of small target pedestrians becomes slower. In order to reduce the parameters and calculation amount of the algorithm and improve the detection speed, the traditional convolution in PR_Resblock_body is replaced with a depthwise separable convolution.

The depthwise separable convolution greatly reduces the parameters of the model by decomposing the traditional convolution into depthwise convolution and point convolution [30]. It is important that the feature extraction capability of the convolution layer is basically unaffected.

The method is mainly divided into two steps:

Step 1. Depth convolution is a 3 × 3 convolution to reduce the number of parameters;

Step 2. Point convolution uses a 1 × 1 convolution kernel to convolve all channels, which reduces the amount of computation.

Figure 4 shows the difference between standard convolution and depthwise separable convolution. Figure 4 (a) is the

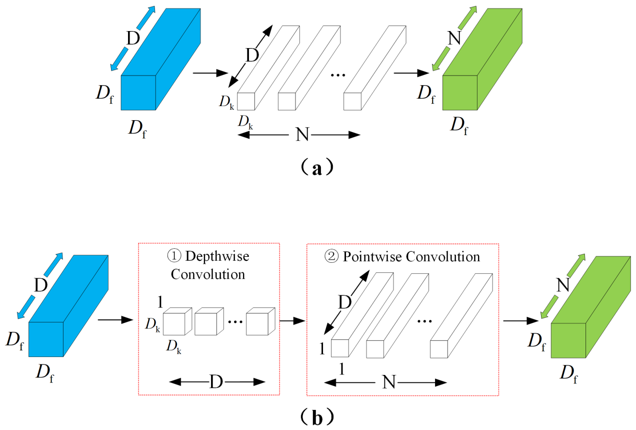


FIGURE 4. Traditional and depthwise separable convolution processes. (a) Traditional convolution, (b) Depthwise separable convolution.

traditional convolution process, and ① and ② in Figure 4 (b) are the depthwise convolution and point convolution in the depthwise separable convolution, respectively.

Suppose feature map size of input is $D_f \times D_f \times D$. The output feature map size is $D_f \times D_f \times N$. The size of the convolution kernel is $D_k \times D_k$. The following is computation comparison of traditional convolution and depthwise separable convolution.

The calculation amount of the traditional convolution C_1 is as follows:

$$C_1 = D_f^2 \times D_k^2 \times D \times N \tag{2}$$

The computation of the depthwise separable convolution C_2 is as follows:

$$C_2 = D_f^2 \times D_k^2 \times D + D_f^2 \times D \times N \tag{3}$$

The ratio of the computational effort of depthwise separable convolution to traditional convolution is as follows:

$$\frac{C_2}{C_1} = \frac{D_f^2 \times D_k^2 \times D + D_f^2 \times D \times N}{D_f^2 \times D_k^2 \times D \times N} = \frac{1}{N} + \frac{1}{D_k^2} \tag{4}$$

It can be seen from the above formula that with the same convolution kernel and channel number, the depthwise separable convolution can greatly reduce the number of parameters and calculations without losing accuracy, and improve the detection speed of the algorithm.

C. CONVOLUTIONAL BLOCK ATTENTION MODULE

Since small target pedestrians occupy very few pixels in the image, the effective features extracted from the backbone network are limited, and there are problems such as target occlusion, resulting in low detection accuracy of small target pedestrians. Therefore adding CBAM after the output features map of the backbone network.

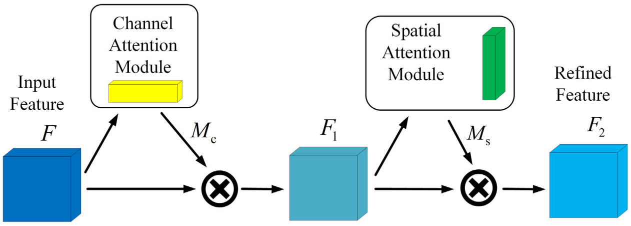


FIGURE 5. Convolutional block attention module.

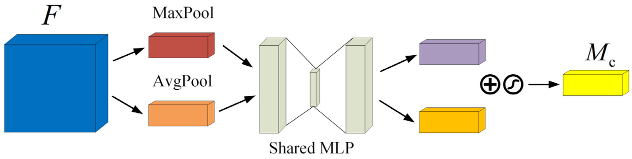


FIGURE 6. Channel attention module.

There are two modules in CBAM, as showed in Figure 5. First, there is the feature map F put in the channel module. The channel attention feature map is generated by the channel relationship, and the feature map F is weighted to obtain the channel feature map F_1. The importance of each channel feature map is obtained through the channel attention module, so that the algorithm pays more attention to the channels with high weight, and suppresses the channels with low weight, which improves the algorithm's ability to extract global features [31]. Then, the relationship between the spatial features is used to supplement the feature information and obtain the spatial feature F_2. The spatial attention module is used to obtain the importance of different regions in the feature map to enhance the algorithm's ability to extract local features [31]. Finally, the input feature map running with weighted operation get the output feature map by the spatial feature map F_2.

In Figure 6. Firstly, feature F of $H \times W \times C$ is input, and H, W and C means the length, width and channels of the feature map. The spatial information of the feature map is aggregated through the global average pooling $AvgPool(F)$ and the maximum pooling $MaxPool(F)$ based on length and width to obtain two $c \times 1 \times 1$ feature information. Then, the feature information is passed through a Multilayer Perceptron (MLP) which has the characteristics of easily assigning weight to feature vectors and keeping the same dimension of the output and input data, respectively. Weights are assigned to each channel of the feature map through MLP to obtain two $c \times 1 \times 1$ feature maps, and the weight coefficient M_c is got through a sigmoid activation function. Finally, the weight coefficients are multiplied by the original feature F to get the feature map F_1, which is the input feature required by the spatial attention module. The formula is shown in formula (5):

$$F_1 = F \times \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{5}$$

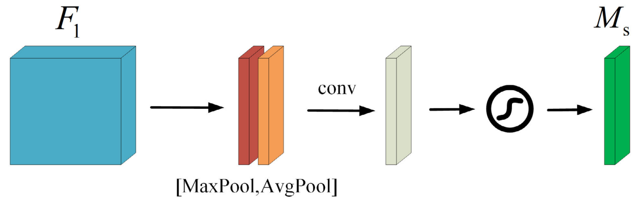


FIGURE 7. Spatial attention module.

where σ is the sigmoid function, which is one of the commonly used neural network activation functions.

In Figure 7, input F_1 into the spatial attention module. After the maximum pooling and average pooling of F_1 respectively, two $H \times W \times 1$ feature information is obtained, and they are spliced together. Then, through a 7×7 convolution layer to get the weight coefficient M_s. Finally, the weight coefficient is multiplied with the feature F_1 to get the feature map F_2. The calculation formula is shown in formula (6):

$$F_2 = F_1 \times \sigma(f^{7\times7}(f^c(MaxPool(F), Avgpool(F)))) \quad (6)$$

where σ is the sigmoid function, f^c means combination, and $f^{7\times7}$ represents filtering using a 7×7 convolution kernel filtrate.

In scenarios such as small target pedestrians which are occluded and pedestrians are too small, the effective information of small target pedestrians that can be extracted is too small. After adding the CBAM, the channel information and spatial information of small target pedestrians are enhanced, which promotes the network to learn more meaningful feature information, and the impact of too little feature information is solved.

D. LOSS

This paper adopts the loss function of YOLOv4 algorithm, which consists of three parts: confidence level loss function, classification loss function and bounding box loss function. The formula is as follows:

$$Loss = Loss_1 + Loss_2 + Loss_3 \quad (7)$$

(1)The calculation formula of the confidence loss function is as follows:

$$\begin{aligned} Loss_1 &= -\sum_i^{S^2}\sum_j^B W_{ij}^{obj}(L_1 + L_2) \\ &\quad - \gamma_{nobj}\sum_i^{S^2}\sum_j^B(1 - W_{ij}^{obj})(L_1 + L_2) \quad (8) \\ L_2 &= \widehat{C}_i^j \log(C_i^j) \quad (9) \\ L_2 &= (1 - \widehat{C}_i^j)\log(1 - C_i^j) \quad (10) \end{aligned}$$

S^2 is the number of grids divided, B is bounding boxes, W_{ij}^{obj} is used to determine the confidence score of the j-th bounding box of the i-th grid, γ_{nobj} is the confidence level of negative samples, $\widehat{C}_i^j$ is the actual confidence level of the presence of positive samples in the i-th grid, and C_i^j is the prediction confidence level of the presence of positive samples in the i-th grid.

(2)The calculation formula of the classification loss function is as follows:

$$\begin{aligned} Loss_2 &= -\sum_i^{S^2}\sum_j^B W_{ij}^{obj}\sum_{c=1}^C(L_3 - L_4) \quad (11) \\ L_3 &= (\widehat{p}_i^j(c))\log(p_i^j(c)) \quad (12) \\ L_4 &= (1 - \widehat{p}_i^j(c))\log(1 - p_i^j(c)) \quad (13) \end{aligned}$$

where $\widehat{p}_i^j(c)$ is the actual probability of category c in the j-th bounding box of the i-th grid, and $p_i^j(c)$ is the prediction probability of category c in the j-th boundary box of the i-th grid.

(3) The bounding box regression loss function adopts the CIoU [32] loss function. CIoU in charge of the distance in target and anchor, the overlap rate, the scale, and the penalty term. In this way, it will make the target box regression more robust. CIoU does not suffer from training instability like IoU [33] and GIoU [34]. In the penalty term of CIoU, the difference between the aspect ratio of the prediction box and the aspect ratio of the target box is included. The CIoU formula is as follows:

$$Loss_3 = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v \quad (14)$$

$\rho^2(b, b^{gt})$ means the Euclidean distance between the center points of the predicted box and the ground-truth box, respectively. c means the diagonal distance of the smallest closure region that can contain both the predicted box and the ground-truth box. The formulas for calculating α and v are as following:

$$\begin{aligned} \alpha &= \frac{v}{1 - IoU + v} \quad (15) \\ v &= \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2 \quad (16) \end{aligned}$$

In formula (16), the parameter ω^{gt} and h^{gt} are the width and height of the real box; ω and h are the width and height of the predicted box.

V. EXPERIMENT

A. DATASET AND IMPLEMENTATION DETAILS

In this experiment, three mixed datasets are used to evaluate the PF_YOLOv4. The datasets are: MS COCO [35], [36], INRIA [37] and self-made datasets.

The MS COCO dataset contains 80 categories of data. The MS COCO dataset includes shopping and vendor targets in squares, pedestrian small targets walking on pedestrian streets, and targets in various scenes. The target scenes are rich and there are a large number of small target pedestrian images. It is suitable for small target pedestrian detection tasks. This paper only uses the single-category data of pedestrians, extracts 1968 images with better features from the dataset, and converts them to the Palcal VOC data format as part of the pedestrian object detection dataset.

The INRIA pedestrian detection dataset includes small target data such as complex human poses and changing lighting conditions. Its photo library is divided into four categories,

TABLE 1. Hardware and software configuration of the experimental platform.

Name	Related configuration
Experimental environment	Windows10
CPU/GHz	Inter Corei7-11800H, 2.3
Memory/(GB)	8
GPU	NVIDIA GeForce GTX 3060, 6G
GPU accelerator	CUDA 11.1, CUDNN 8.0
Deep learning framework	Tensorflow 2.4

namely, only cars, only people, people with cars, and people without cars. The training set and the test set constitute the data set. The training set contains positive and negative samples, including 614 positive samples and 1218 negative samples, and 288 positive samples and 453 negative samples in the test set. This paper extracts 564 positive samples from the INRIA pedestrian detection dataset, and converts its data format to Pascal VOC.

The self-made dataset is 862 pedestrian pictures taken from real life, including rainy day scenes, traffic light scenes and complex scenes under night street light scenes. There are many small targets in these pictures. According to the labeling of Pascal VOC, the Labelling labeling tool is used in the training process to generate a label.xml for each picture in the xml file.

Combining all the above data sets, the number of experimental samples is 3394. Before the network model is trained, the data set is divided into training set and test set data, the ratio is 75% and 25%, and then the four-fold crossover method is used to decompose the training set and test set. The average of the four experimental results was taken as the experimental result.

The experimental platform configuration in this paper is shown in Table 1.

Reasonable setting of network parameters can greatly improve algorithm training and detection results. First load the pre-training algorithm, then set batch = 64, subdivisions = 16. Each iteration will randomly select 64 samples from all training sets for training, and these 64 samples are divided into 16 groups equally and sent to the network for training. The parameters used throughout the training process are as follows: momentum is 0.9, weight decay coefficient is 0.001, and initial learning rate is 0.0001, maximum number of iterations of 20000, and learning at 13,000 and 17,000 iterations rate is reduced by a factor of 10. On this basis, the pedestrian dataset is used for algorithm training and log files are saved.

B. EVALUATION CRITERIA

In this paper, mAP value is selected as the test result accuracy index. Recall is the evaluation index of recall rate, IOU is the positioning accuracy evaluation index, and FPS is the real-time detection speed index.

$$mAP = \frac{\sum_{i=0}^{k} AP_i}{h} \tag{17}$$

TABLE 2. Ablation study. PR stands for soft thresholding module, D stands for depthwise separable convolution module.

No.	PR	D	CBAM	mAP(%)	FPS(s)
YOLOv4				93.27	57
1	√			94.91	49
2		√		92.85	73
3			√	94.31	52
4	√	√		94.87	59
5	√		√	95.39	45
6		√	√	94.95	65
PF_YOLOv4	√	√	√	95.62	58

The definition of mAP is shown in formula (17), AP represents the accuracy rate corresponding to each target, and h represents the total number of target object categories.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

The definition of Recall is shown in formula (18), where True Positives (TP) is the number of positive samples correctly marked as positive samples, False Negatives (FN) is belong to the quality of positive sample wrongly identified as negatives samples.

$$IOU = \frac{S_{A\cap B}}{S_A + S_B - S_{A\cap B}} \tag{19}$$

Intersection over Union (IOU) is defined as showed in formula (19). S_A Represents the area of the box A, S_B represents the area of box B, and $S_{A\cap B}$ represents the area of the intersection of box A and box B.

FPS is the definition in the image field, which is about the number of frames per second transmitted by the screen. FPS value reflects the amount of information saved and displayed in dynamic video. The higher the FPS, the more frames per second, and the smoother the displayed video.

C. ABLATION STUDY

To testify the performance of proposed algorithm and analyze each novel method performance, six groups of ablation experiments were designed on the basis of YOLOv4. Each group of experiments used the same hyperparameters and training techniques. The results are shown in Table 2 of show:

As shown in Table 2, after adding the soft thresholding module, the detection accuracy is improved by 1.64%, indicating that the soft thresholding module effectively reduces the interference of light and other noise on the detection of small target pedestrians and improves the detection accuracy. The depthwise separable convolution module is added, and the detection speed is increased by 16 FPS. The depthwise separable convolution decrease the amount of parameters and computation of the model, and improves the detection speed. After adding CBAM, the detection accuracy is improved by 1.04%, indicating that CBAM can improve the feature extraction ability of the algorithm for small target pedestrians and improve the detection accuracy. After adding the soft thresholding module and the depthwise separable convolution module, the detection accuracy is increased by 1.6%, and

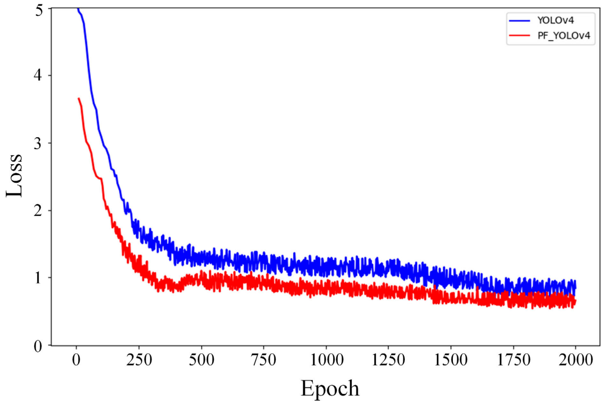


FIGURE 8. Loss curve.

the detection speed is increased by 10 FPS compared with adding the soft thresholding module alone. It shows that the soft thresholding module and the depthwise separable convolution module are added to the algorithm, which improves the detection accuracy and speed of the algorithm, and proves the effectiveness of the two modules. After adding the soft thresholding module and CBAM, the detection accuracy is increased by 2.12%. Compared with adding a single module, adding the soft thresholding module and CBAM at the same time achieves better results. After adding the depthwise separable convolution module and CBAM, the detection accuracy is increased by 1.68%, and the detection speed is increased by 13 FPS compared with adding CBAM alone, indicating that the depthwise separable convolution module and CBAM are added to the algorithm at the same time, not only can enhance the characteristics of small target pedestrians the extraction capability can also improve the detection speed, indicating that both modules are effective.

In summary, adding soft threshold module or CBAM improves the detection accuracy, but increases the complexity of the algorithm. The introduction of depthwise separable convolution can effectively reduce the complexity of the network, greatly reduce the amount of computation, and achieve a balance between speed and accuracy.

D. DETECTION RESULTS

In the training, the training effect of the algorithm is judged by the change of loss value. Figure 8 demonstrates the change of the loss value of the PF_YOLOv4 algorithm and the YOLOv4 algorithm during the training process of 2000 Epochs.

In Figure 8, loss values of both algorithms decrease rapidly in the first 500 iterations, and the improved PF_YOLOv4 algorithm decreases faster than the YOLOv4 algorithm. After 420 iterations of YOLOv4 algorithm, the loss value dropped from 5 at the beginning to around 1.4, and then began to slowly decrease, and finally after 1600 iterations, the loss value tended to stabilize. PF_YOLOv4 algorithm after 300 iterations, the loss value dropped rapidly from 4.6 to around 1, and then began to decline steadily, and stabilized after about 1300 iterations. The above analysis that in the training process, the loss value of the improved PF_YOLOv4 algorithm decreases more rapidly than that of the YOLOv4 algorithm, and the loss value after stabilization is also smaller, indicating that the PF_YOLOv4 algorithm is more stable and converges faster.

TABLE 3. Performance evaluation of different algorithms on test data set.

Network model	mAP(%)	Recall(%)	IOU(%)	FPS(s)
Faster R-CNN	75.68	76	76.59	46
YOLOv3	85.95	79	85.14	51
YOLOv4	93.27	83	91.19	57
YOLOv5	92.13	81	90.21	65
PF_YOLOv4	95.62	89	93.36	58

To evaluate the detection accuracy of the algorithm more rigorously, the PF_YOLOv4 algorithm is compared with the classic target detection algorithm, and the results are shown in Table 3.

As showed in Table 3, Faster R-CNN relies on generating candidate boxes, and the detection accuracy reaches 75.68%, but due to its complex structure, the detection speed is low. The PF_YOLOv4 algorithm adds soft thresholding module and CBAM, which is able to modify the detection accuracy of the algorithm. The mAP value is 19.94% higher than that of Faster R-CNN. The added depth is separable convolution module improves the algorithm detection speed, which is 12 FPS faster than Faster R-CNN. The YOLOv3 algorithm is the representative of the end-to-end detection algorithm. It removes the stage of generating candidate regions and directly obtains the object position coordinates value and category probability, which is able to greatly increase the target detection speed. Compared with the YOLOv3 algorithm, the mAP of the PF_YOLOv4 algorithm is increased by 9.67%, and the detection speed is increased by 9 FPS. Compared with the YOLOv4 algorithm, the PF_YOLOv4 algorithm improves the detection accuracy by 2.35% and the detection speed by 1 FPS. Although the detection speed of the YOLOv5 algorithm is faster than that of the PF_YOLOv4 algorithm, the detection accuracy of small target pedestrians on the same dataset is lower than that of the YOLOv4 algorithm and the PF_YOLOv4 algorithm. Compared with the YOLOv5 algorithm, the mPA of the PF_YOLOv4 algorithm is improved by 3.49%, and it has an excellent detection effect in the detection of small target pedestrians. From the above analysis, it can be seen that the PF_YOLOv4 algorithm proposed in this paper has better performance in small target pedestrian detection, higher detection accuracy, and also meet real-time requirements.

YOLOv4 algorithm, YOLOv5 algorithm and PF_YOLOv4 algorithm all used mosaic for data enhancement in the data preprocessing process, and tested on the same dataset. Part of the detection effect diagram is shown in Figure 9.

According to the above sets of test results, it can be found that the YOLOv4 algorithm and YOLOv5 algorithm misses the detection seriously in the dark environment where the small target pedestrian is located. The PF_YOLOv4 adds a

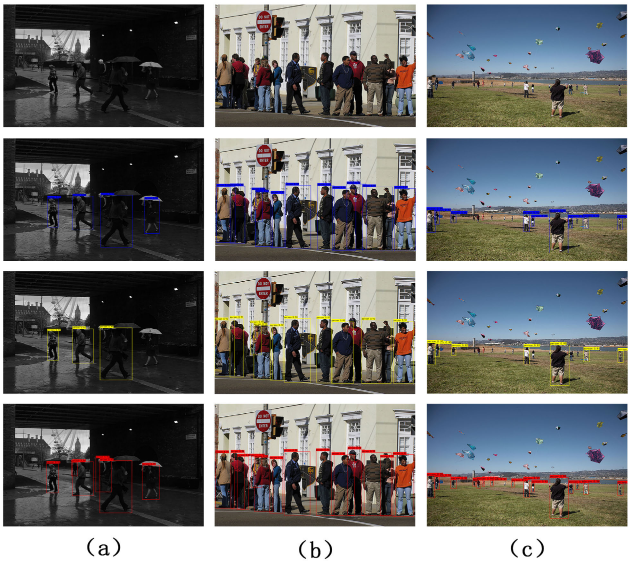


FIGURE 9. YOLOv4, YOLOv5 and PF_YOLOv4 detection results. The second row is the YOLOv4 detection result, the third row is the YOLOv5 detection result, and the fourth row is the PF_YOLOv4 detection result. (a)Small target pedestrians in poor light environment. (b)Small target pedestrians are occluded. (c)Small target pedestrians are too small.

soft thresholding module, which greatly reduces the impact of noise such as light on the detection of small target pedestrians. When there are a large number of small target pedestrians in the image and the targets overlap, the YOLOv4 algorithm and YOLOv5 algorithm has low recognition accuracy and obviously missed detection, while the PF_YOLOv4 algorithm performs well. When the pedestrian target is small in the image, the PF_YOLOv4 algorithm can better detect small pedestrians due to the addition of CBAM, and the detection accuracy is higher than the YOLOv4 algorithm and YOLOv5 algorithm.

The above detection results show that the PF_YOLOv4 algorithm has stronger noise reduction ability and better anti-interference ability, and has good performance for small target pedestrian detection in various complex situations, which are significantly improved compared to YOLOv4 algorithm and YOLOv5 algorithm.

VI. CONCLUSION

This paper proposes the PF_YOLOv4 algorithm for small target pedestrian detection by adding a soft thresholding module, a depthwise separable convolution module and a CBAM. The method can reduce noise interference such as light, and achieves satisfactory performance in average precision and missed detection rate for pedestrians with fewer features and small scales. Datasets such as MS COCO and INRIA are employed to train, validate and test. The experimental results show that the mAP of the PF_YOLOv4 algorithm is increased by 19.94% and the speed is increased by 12 FPS compared with the two-stage detection algorithm Faster R-CNN. Compared with YOLOv4, the mAP value is increased 2.35% and the detection speed is also improved. PF_YOLOv4 algorithm has achieved good performance in reducing noise interference, occlusion of small target pedestrians and smaller small target pedestrians. It demonstrates that the PF_YOLOv4 has excellent detection performance and robustness in small target pedestrian detection.

REFERENCES

[1] B. Bosquet, M. Mucientes, and V. M. Brea, "STDNet: Exploiting high resolution feature maps for small object detection," *Eng. Appl. Artif. Intell.*, vol. 91, May 2020, Art. no. 103615.

[2] D. Zeng, F. Zhao, S. Ge, and W. Shen, "Fast cascade face detection with pyramid network," *Pattern Recognit. Lett.*, vol. 119, pp. 180–186, Mar. 2019.

[3] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Finding tiny faces in the wild with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 21–30.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

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

[6] 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.

[7] M. Ansari and K. Lodi, "A survey of recent trends in two-stage object detection methods," in *Proc. Int. Conf. Renewal Power (ICRP)*. Singapore: Springer, 2021, pp. 669–677.

[8] Y. Zhang, X. Li, F. Wang, B. Wei, and L. Li, "A comprehensive review of one-stage networks for object detection," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2021, pp. 1–6.

[9] X. Yang, J. Zhao, H. Zhang, C. Dai, L. Zhao, Z. Ji, and I. Ganchev, "Remote sensing image detection based on YOLOv4 improvements," *IEEE Access*, vol. 10, pp. 95527–95538, 2022.

[10] 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.*, Jun. 2014, pp. 580–587.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2014.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[14] K. C. Lai, J. Zhao, D. J. Liu, X. N. Huang, and L. Wang, "Research on pedestrian detection using optimized mask R-CNN algorithm in low-light road environment," *J. Phys., Conf. Ser.*, vol. 1777, no. 1, Feb. 2021, Art. no. 012057.

[15] 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.* Springer, 2016, pp. 21–37.

[16] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[17] Q. Zhang, "A novel ResNet101 model based on dense dilated convolution for image classification," *Social Netw. Appl. Sci.*, vol. 4, no. 1, pp. 1–13, Jan. 2022.

[18] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*.

[19] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: Fast detection for small objects," *Proc. SPIE*, vol. 10615, pp. 381–388, Apr. 2018.

[20] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.

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

[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[25] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.

[26] W. Boyuan and W. Muqing, "Study on pedestrian detection based on an improved YOLOv4 algorithm," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1198–1202.

[27] Z. Cao, H. Yang, J. Zhao, S. Guo, and L. Li, "Attention fusion for one-stage multispectral pedestrian detection," *Sensors*, vol. 21, no. 12, p. 4184, Jun. 2021.

[28] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[29] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[32] 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.*, 2020, vol. 34, no. 7, pp. 12993–13000.

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

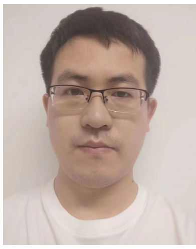
[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[35] Y. Zhao, F. Shi, M. Zhao, W. Zhang, and S. Chen, "Detecting small scale pedestrians and anthropomorphic negative samples based on light-field imaging," *IEEE Access*, vol. 8, pp. 105082–105093, 2020.

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

KAIHUI LI was born in Shandong, China, in 1995. He is currently pursuing the master's degree with the Qilu University of Technology, Jinan, China. His research interests include image processing, artificial intelligence, and machine learning.



YUAN ZHUANG was born in 1991. He received the master's degree. He is currently an Engineer. His main research interest includes high-performance computing.

JINLING LAI was born in 1998. He received the master's degree in computer science and technology (artificial intelligence) from the Qilu University of Technology, in 2021. His research interests include deep learning, artificial intelligence, and object detection.

YUNHUI ZENG was born in 1975. He received the Ph.D. degree. He is currently a professor. His main research interests include numerical simulation and high-performance computing. He is a member of the China Computer Federation (CCF).

• • •