ITC 2/53
Information Technology and Control
Vol. 53 / No. 2 / 2024
pp. 390-407
DOI 10.5755/j01.itc.53.2.35569

YOLOv7-PD: Incorporating DE-ELAN and NWD-CIoU for Advanced Pedestrian Detection Method

Received 2023/11/09 | Accepted after revision 2024/03/10

HOW TO CITE: He, Y., Wan, L. (2024). YOLOv7-PD: Incorporating DE-ELAN and NWD-CIoU for Advanced Pedestrian Detection Method. *Information Technology and Control*, 53(2), 390-407. https://doi.org/10.5755/j01.itc.53.2.35569

YOLOv7-PD: Incorporating DE-ELAN and NWD-CIoU for Advanced Pedestrian Detection Method

Yu He
State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou, China; e-mail: 1542997906@qq.com

Liang Wan
State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou, China; e-mail: lwan@gzu.edu.cn

Corresponding author: lwan@gzu.edu.cn(LW)

In the pedestrian detection task, the excessive depth of the convolutional network in YOLOv7 results in an abundance of background feature information, thereby posing challenges for the model to accurately locate and detect pedestrians, particularly in small-scale or heavily occluded scenarios. To handle this problem, we propose a pedestrian detection model called YOLOv7-PD, to strengthen the accuracy of detecting small-scale pedestrians and occluded pedestrians. First of all, we propose an improved module called DE-ELAN, an improvement on the existing E-ELAN module, which is based on Omni-Dimensional Dynamic Convolution (ODConv). This module leverages four complementary attention types to enhance feature extraction, capturing rich contextual information. Then, we propose a lightweight receptive field enhancement module called light-REFM, which constructs a pyramid structure and acquires fine-grained multi-scale information through dilated convolutions of different sizes. Finally, we propose an improved regression loss function based on the Normalized Wasserstein Distance (NWD) that combines NWD with Complete-IoU (CIoU), enabling precise position and feature capture for small targets. On the Citypersons dataset, YOLOv7-PD outperforms YOLOv7, improving the average precision (AP) by 7% and reducing the miss rate by 2.58%. Experiments on three challenging pedestrian detection datasets demonstrate a balance between precision and speed, achieving excellent performance.

KEYWORDS: Pedestrian detection, YOLOv7, ODConv, CIoU, NWD.

1. Introduction

Pedestrian detection is a crucial detection task within the domain of computer vision and has a broad range of applications in the area of autonomous cars [26, 41]. Pedestrian detection is a method of automatically identifying the location and dimensions of pedestrians within a picture or video using computer vision methods. Pedestrians are one of the main participants in the road traffic environment, they are not fixed like rigid objects but have changeable shapes. Pedestrians can have different postures and different appearances, and the background environment is also diverse, which makes the difficulty of pedestrian detection greatly increased. At present, pedestrian detection faces the challenges associated with precise recognition and the precise location of small-scale object targets and occluded object targets [14, 23, 30, 36]. Over the past few years, YOLOv7 [44] has received widespread acknowledgment due to its outstanding detection speed and high-precision object detection. However, the YOLOv7 continues to face two problems with pedestrian detection. 1) Excessively deep convolutional networks in YOLOv7 can lead to an overabundance of background feature information, making it challenging for the model to precisely locate and detect pedestrians in situations involving small scales and severe occlusions. 2) For pedestrians of different sizes, the sensitivity of IoU varies greatly. For small target pedestrians, a slight position deviation results in a significant reduction in IoU.

To solve this issue, we present a pedestrian detection model referred to as YOLOv7-PD. We propose an improved E-ELAN module (DE-ELAN), which enhances feature extraction performance and contributes to a better capture of rich contextual information. Then, we propose a lightweight-receptive field enhancement module (light-REFM) to obtain fine-grained multiscale information. Last, we propose an upgraded method for regression loss function to boost the small target pedestrian's detection accuracy.

This paper makes an important contribution to solving the above problems, and the summary is as follows:

1 We propose an improved E-ELAN module (DE-ELAN) that leverages four complementary types of attention and progressively applies attention to various dimensions of the convolutional operation. This significantly improves feature extraction performance, resulting in a better capture of rich contextual information.

2 We propose a lightweight-receptive field enhancement module (light-REFM) designed to construct a pyramid structure and capture fine-grained multi-scale spatial information of various channels using dilated convolution of different sizes.

3 We propose an improved regression loss function method, we introduced the Normalized Wasserstein Distance (NWD) [45] into the regression loss function and combined it with the CIoU method [9], which alleviates the constraints of CIoU in small-target detection.

2. Related Work

In this section, we explore a variety of issues concerning the content of this paper, containing pedestrian detection, attention mechanism, and feature pyramid.

2.1. Pedestrian Detection

Pedestrian detection mainly determines whether there is a pedestrian target through the given static image or dynamic video by the computer [26]. If there is a pedestrian, the bounding box is employed to label the specific location of the pedestrian and give the confidence score [6,9]. With the vigorous advancement of artificial intelligence technology, pedestrian detection has shown a broad spectrum of application scenarios in the arena of automatic driving, human-computer interaction, intelligent video surveillance, and urban street view [12, 29]. In the early, pedestrian detection methods for manual feature extraction. Dalal et al. [11] proposed the Histogram of Oriented Gradients (HOG) method, utilizing edge direction and intensity information to characterize the overall visual representation of pedestrians. Generally speaking, manual feature extraction methods have certain advantages in the realm of pedestrian detection, however, the extraction steps are cumbersome. Over the years, deep learning techniques have achieved significant breakthroughs in computer vision tasks [53, 54, 55, 56], especially in pedestrian detection [38, 39]. Pedestrian detection algorithms that

utilize deep learning approaches have improved the accuracy of detection algorithms, which can be sorted as either two-stage detection approaches or one-stage detection approaches [52]. The two-stage detection approaches have high accuracy, however the speed of detection is significantly sluggish. The Faster R-CNN proposed by Girshick et al. [18], utilizes the Region Proposal Network (RPN) network to directly produce region proposals. The SA-FastRCNN proposed by Li et al. [25], incorporates a unified structure by combining two parallel large subnetworks into an integrated architecture, supplying confidence scores for different classes and object bounding box regression for various sizes. The one-stage detection approaches have advantages in terms of speed, simplicity, and real-time performance, rendering them appropriate for many computer vision applications, particularly those with high-speed processing requirements. Liu et al. [33] proposed SSD, which is directly detected using CNN. Bochkovskiy et al. [44] proposed YOLOv7. The backbone of the YOLOv7 model is primarily constructed with convolutional layers, E-ELAN modules, and MPConv modules. In particular, the E-ELAN module, building upon the original ELAN, modifies the computation block while preserving the transitional layer architecture of the initial ELAN design. It boosts the network's learning capacity by incorporating the ideas of expanding, shuffling, and merging cardinality without disrupting the existing gradient path.

2.2. Attention Mechanism

The attention mechanism is a crucial technology in deep learning that aims to enable models to distribute different attention or weights to different parts of input data, allowing them to concentrate on important information during data processing, thereby improving model performance. The key point of the attention mechanism is similar to the attention focus in human perception processes, enabling adaptive concentration on different parts when processing sequences, images, or other types of data [2, 37]. Hu et al. [22] proposed SENet, utilizing a breakthrough channel attention module named "Squeeze and Excitation" (SE) to leverage the interdependencies between convolutional feature channels. Taking inspiration from the SE block, the ECA block is proposed by Wang et al. [46], offering a more effective channel attention design by taking the place of the initial fully connected layer of SE with cost-effective 1D convolutions and reducing an extensive set of parameters. Woo et al. [47] proposed a lightweight module of attention called CBAM, combining the channel attention module with the spatial attention module. Misra et al. [35] proposed an attention module that has three branches called Triplet Attention, which is conditioned on features that rotate along three different dimensions. The attention mechanism aids the model in concentrating pivotal regions in the image, thereby enhancing the performance of object detection and localization tasks. By incorporating attention within convolutional layers, the network can autonomously learn to selectively emphasize critical objects or areas.

2.3. Feature Pyramid

The scale variation in complex road scenes significantly affects the accuracy of pedestrian detection. The feature pyramid is a multi-scale feature representation method used in computer vision tasks. Its main goal is to process information at different scales and capture both local details and the global context of an object in an image [27]. Inspired by SPP [19], Chen et al. [5] proposed the ASPP module, employing multiple parallel dilated convolution layers with varying sampling rates. Features gathered at various sampling rates are subsequently treated in separate branches and merged to produce the ultimate result. The module creates convolution kernels with varying receptive fields by various dilation rates, aiming to capture multi-scale object information. The RFB module was proposed by Songtao Liu et al [31]. This module simulates a range of observations similar to human vision to augment the network's feature extraction efficiency. It combines different receptive fields by using different convolutional kernels and different step sizes, uses 1×1 convolutions for dimensionality reduction, and ultimately constructs a hybrid superposition of varying receptive fields.

3. The Proposed Method

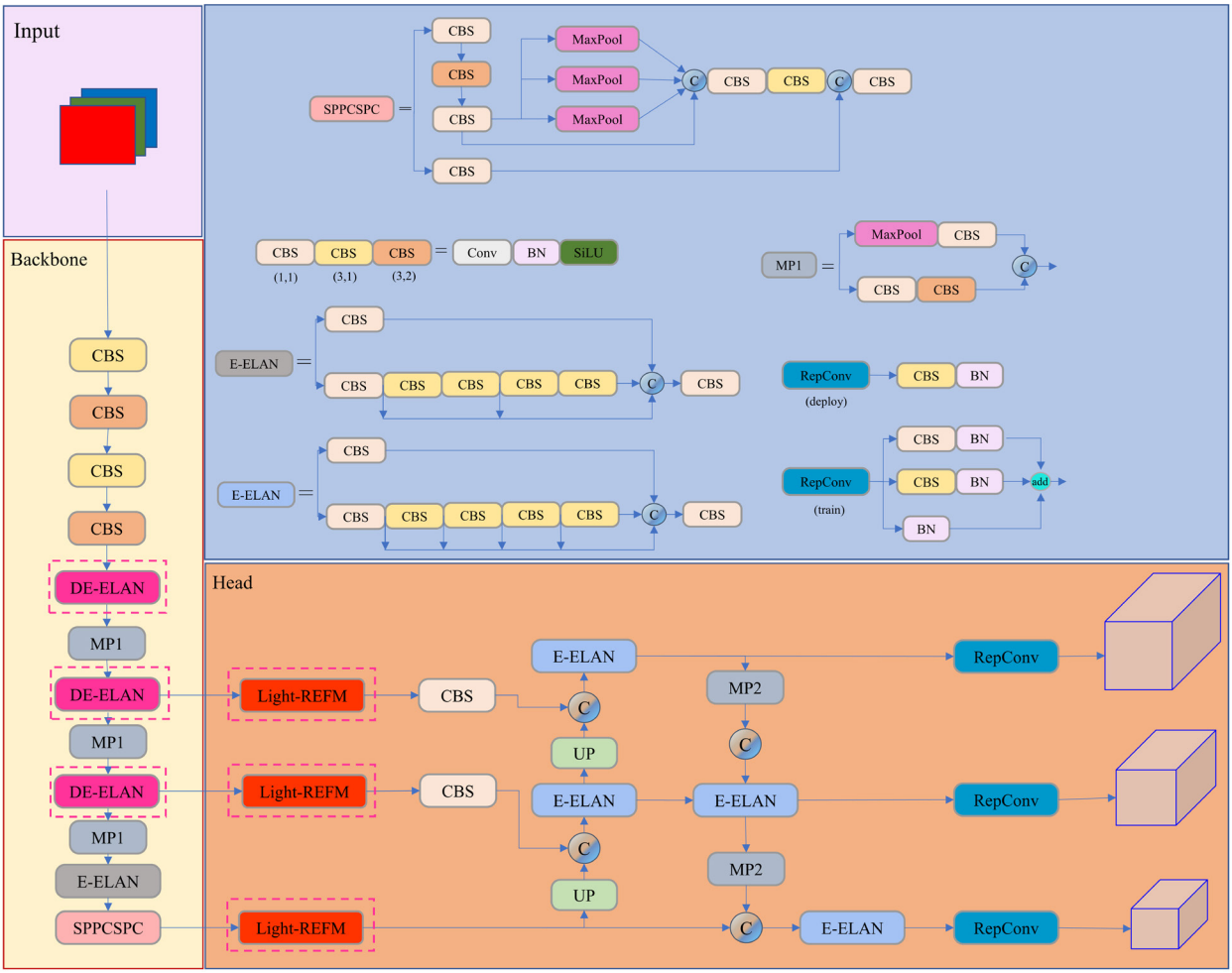
Due to differences in the shape, size, and other attributes of pedestrians in complex road environments, as well as possible occlusion, the original YOLOv7 network may not be able to accurately locate and detect pedestrians. Therefore, we propose the YOLOv7-

PD network. The network initially enhances the backbone of the YOLOv7, proposing the DE-ELAN module based on Omni-dimensional Dynamic Convolution that optimizes the original E-ELAN module. This aims to boost the network's learning capacity for pedestrian targets with different shapes, sizes, and occlusion levels to capture richer feature information. Secondly, in the Head network, we propose the lightweight receptive field enhancement module (light-REFM), which enhances the precision of multi-scale pedestrian detection and identification. Thirdly, the Normalized Wasserstein Distance (NWD) loss is integrated into the regression loss function and combined with the CIoU method to boost the effectiveness of YOLOv7 in the small target pedestrian detection task. Figure 1 illustrates the overall architecture of YOLOv7-PD.

3.1. E-ELAN Module Based on Omni-dimensional Dynamic Convolution

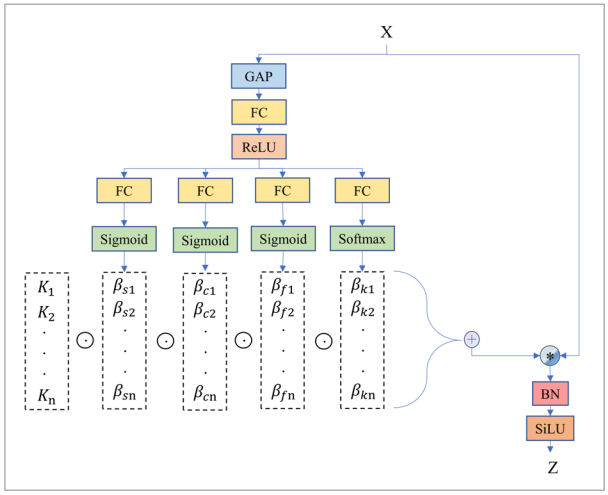
In pedestrian detection, the posture, shape, and size of pedestrians may vary due to different shooting angles and distances. Thus, by incorporating full-dimensional dynamic convolutions into the backbone network's E-ELAN module and gradually introducing different attention mechanisms in various dimensions (such as position, channel, filters, and convolution kernels) into the convolution operation, the convolution pro-

Figure 1
The overall architecture of YOLOv7-PD



cess can better adapt to differences across various aspects of input data. Consequently, it can more effectively capture rich contextual information for pedestrians of different scales. Therefore, we designed an improved E-ELAN module (DE-ELAN) to replace the last CBS module in the E-ELAN in YOLOv7's feature extraction network with an improved OCBS module. The OCBS module based on Omni-dimensional Dynamic Convolution is displayed in Figure 2.

Figure 2
The structure of OCBS

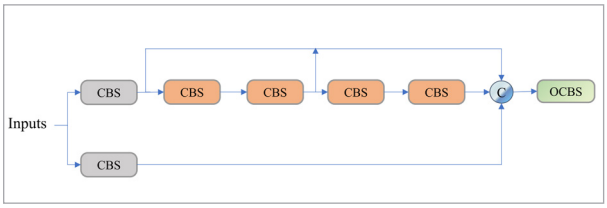


In the OCBS module, we introduce ODConv (Omni-dimensional Dynamic Convolution) [24], which uses a parallel strategy to employ a multi-dimensional attention mechanism along four dimensions in the kernel space for learning more flexible and complementary attention. It simultaneously considers account dynamics across dimensions including spatial, input channels, output channels, etc., to capture rich contextual information. This multi-dimensional manipulation can improve the model's capability to analyze data, enhance its perception of different features, and contribute to better performance on complex tasks. The OCBS can be described in the following:

$$Z=\left(\beta_{k1}\odot\beta_{f1}\odot\beta_{c1}\odot\beta_{s1}\odot K_1+\cdots+\beta_{kn}\odot\beta_{fn}\odot\beta_{cn}\odot\beta_{sn}\odot K_n\right)*x, \quad (1)$$

where $\beta_{ki}\epsilon R$ represents the attention scalar for the convolutional kernel K_i; $\beta_{si}\epsilon R^{k\times k}$, $\beta_{ci}\epsilon R^{c_{in}}$ and $\beta_{fi}\epsilon R^{c_{out}}$ represent attention points that are applied across the spatial, input channel, and output channel dimensions respectively. These points are calculated across the spatial, input channel, and output channel dimensions of the convolution kernel W_i. $\odot$ represents the multiplication of various dimensions along the kernel space. β_{si}, β_{ci}, β_{fi} and β_{ki} are calculated using a multi-head attention model $\varpi_i(x)$. Traditional convolution still plays a certain role in pedestrian detection, especially for pedestrians with moderate sizes and relatively normal postures. However, its adaptability is limited when dealing with pedestrians with irregular shapes and significant variations in posture. Therefore, we designed the DE-ELAN module to boost the ability to capture details of small-sized pedestrians. The DE-ELAN module is depicted in Figure 3.

Figure 3
The structure of DE-ELAN

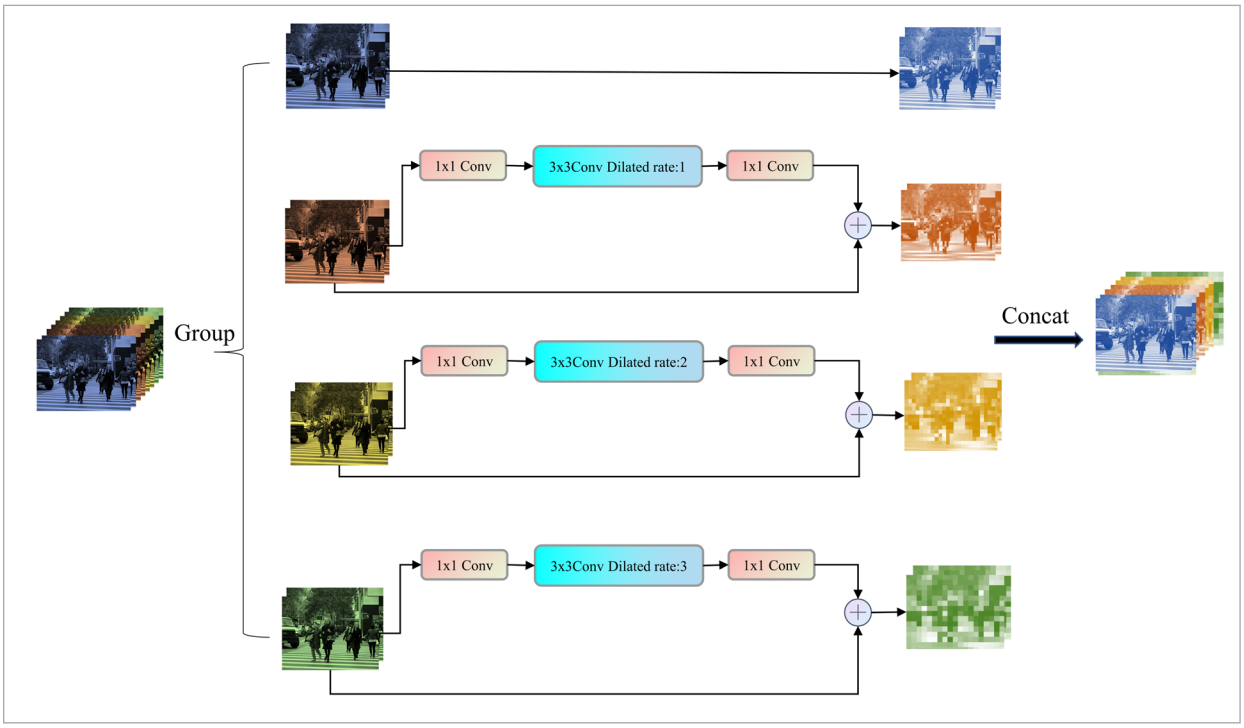


3.2. Lightweight-Receptive Field Enhancement Module

Because different sizes of receptive fields imply different abilities to capture remote dependencies, cover more of the surrounding area, and capture richer contextual information. We propose a lightweight receptive field enhancement module called Light-REFM, which captures the spatial details of the input feature maps on multiple scales through parallel dilated convolution to obtain feature maps that should contain contextual information. The composition of the module is displayed in Figure 4.

Zhang et al. [48] proposed group convolution as a technique to extract features across various scales. The method enabling the processing of channels in groups to capture multi-scale information may suffer from reduced parameter efficiency as the convolution kernel dimensions grow. The erode operation is often used to improve detection results by removing small false detection areas in the image due to factors such as noise or background interference. However, the erode operation may eliminate important feature details when the pedestrian part is not clearly visible or resembles the background. For small target pedes-

Figure 4
The structure of Light-REFM



trians, the erode operation may reduce their effective scale, thereby limiting the flexibility of the model and accuracy in handling pedestrians of different scales. The dilated convolution keeps the size of the feature map while increasing the spatial resolution of the feature map, which is beneficial for preserving detailed features. Therefore, we employ dilated convolutions with different dilation rates to capture feature information across different scales. In the Light-REFM module, the input feature map f is first partitioned into four parts based on the arrangement order of feature channels, resulting in $[F_0, F_1, F_2, F_3]$ grouped by the channel dimension. The quantity of channels in each grouping section is C`=C/4, where C should be a multiple of 4. After each grouping, the resulting feature map is represented as $F_j \in R^{C` \times H \times W}$, where j =0,1,2,3. We retain the original features in the first feature channel group F_0 without introducing additional dilated convolutions, ensuring that the network can capture low-resolution fine details and thus preserve the original information. In the remaining three feature channel groups F_1, F_2, and F_3. We employed dilated convolutions with different dilation rates to different features at varying scales. Therefore, expanding the receptive field and enabling the network to more effectively capture features of varying scales within the image. Finally, by concatenating the outputs of these four distinct branches, the model merges feature information from different scales and resolutions. This helps the model achieve a stronger representational capacity, thus enabling it to better capture multi-scale information and contextual relationships. Simultaneously, it controls the number of parameters, avoiding excessive computational burden. The entire multi-scale feature map F is:

$$F = concat\left(F_j\right), j = 0,1,2,3, \tag{2}$$

where $F \in R^{C` \times H \times W}$ denotes the resulting multi-scale feature map, and contact denotes the act of concatenation.

3.3. Regression Loss Function Design

In YOLOv7, the CIoU method is used as an evaluation metric to compute the overlap region between the predicted box and the ground truth box. Compared to the traditional IoU (Intersection over Union) mea-

surement method, the CIoU method takes into account the normalized differences in the center-point distance, width, and height of the bounding box, making it more robust regarding the bounding box position and size [57]. However, in pedestrian detection, there are pedestrians with different target sizes. For small target pedestrians, the bounding box is relatively small, and even a small positional deviation can lead to a notable change in the IoU value. Although CIoU loss enhances the stability of IoU loss by considering the distance between the centers of bounding boxes and the aspect ratio, these improvements may still not be enough to completely overcome the instability of the IoU. For small targets, a slight offset in the bounding box can lead to a significant increase in the loss value, making the model overly sensitive to the position of small targets, thereby affecting the stability of training and the final performance of the model. In addition, sensitivity to size changes in CIoU loss is particularly prominent on small targets. This may lead the model to be overly sensitive to size changes when dealing with small targets while ignoring other important features such as appearance and contextual information. The NWD is a novel approach for small object detection [45], which uses a novel metric based on Wasserstein distance to calculate bounding box similarity, substituting the conventional IoU measurement method. First, this method models bounding boxes by utilizing two-dimensional Gaussian distributions. Then, Wasserstein distances are used to calculate the resemblance between the corresponding Gaussian distributions. Because this distance can calculate the similarity of distributions even when overlaps are ignored. Furthermore, this method is insensitive to multi-scale objects, making it well-suited for measuring small objects.

We integrate the NWD into the regression loss function alongside the CIoU method to harness the strengths of both in pedestrian detection. CIoU excels in measuring spatial location and overlaps, making it particularly effective for medium to large-scale pedestrians where precision in bounding box alignment is crucial. On the other hand, NWD demonstrates remarkable robustness in handling small-scale pedestrians by effectively capturing the distributional characteristics of the target without being overly sensitive to scale. Its strength lies in recognizing the subtle distributional nuances of small targets, often overlooked by conventional methods. Therefore, by combining CIoU's precision in spatial alignment with NWD's sensitivity to distributional attributes, we significantly boost the model's robustness and accuracy across multiple scales. This fusion allows for a more comprehensive and nuanced capture of the position and features of pedestrians, significantly boosting detection performance, especially in scenarios where targets vary widely in size. The formulation of this regression loss function is expressed as:

$$L_{reg} = (1-\gamma) L_{NWD} + \gamma L_{CIoU}, \tag{3}$$

where L_{reg} is the regression loss function, γ is a weighting factor used to equilibrium the contribution of NWD and CIoU in the regression loss. The value of γ is usually between 0 and 1. It has been verified that the optimal result is achieved when $\gamma = 0.7$. The CIoU loss function is specified as:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v, \tag{4}$$

where IoU is used to assess the level of overlap between the predicted bounding box b and the ground truth bounding box b_{gt}. $\rho^2(b,b_{gt})$ calculates the Euclidean distance between the center point of the predicted box and the ground truth box. c denotes the diagonal distance of the smallest enclosed area that can encompass both the prediction box and the ground truth box. α serves as the trade-off parameter, which employs the importance of the center distance in the loss function. v serves as the function that quantifies the aspect ratio metric.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2. \tag{5}$$

In Equation (5), w and h represent the height and width of the prediction box. w_{gt} and h_{gt} represent the height and width of the ground truth box. The NWD loss function is specified as:

$$L_{NWD} = 1 - NWD(\mathrm{N}_a, \mathrm{N}_b), \tag{6}$$

$$NWD(\mathrm{N}_a, \mathrm{N}_b) = \exp\left(-\frac{\sqrt{w_2^2\left(\mathrm{N}_a, \mathrm{N}_b\right)}}{C}\right), \tag{7}$$

$$W_2^2(\mathrm{N}_a, \mathrm{N}_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2 \tag{8}$$

where c is a constant that has a strong correlation with the dataset, $W2\ 2(\mathcal{N}_a, \mathcal{N}_b)$ is a distance metric. $\mathcal{N}_a, \mathcal{N}_b$ is a Gaussian distribution modeled by boundary boxes A(cx_a, cy_a, w_a, h_a) and B(cx_b, cy_b, w_b, h_b).

4. Experiments

In this section, we utilize three various datasets from CityPersons [50], Caltech [13], and CrowdHuman [40] for training, validation, and testing of our model. We conducted ablation experiments to compute the performance of our proposed method. Finally, we compared our method with the most advanced pedestrian detection methods.

4.1. Dataset and Evaluation Indicators

The Citypersons dataset is a portion of the CityScapes dataset [10], containing annotations for pedestrian objects in images captured near roads using onboard vehicle cameras. This dataset includes street scenes captured in 27 different cities, featuring diverse pedestrian samples. It is partitioned into a training set including 2975 images, 500 images for a validation set, and 1525 images for a test set. The Citypersons dataset is displayed in Figure 5.

CrowdHuman dataset is a dataset for pedestrian detection released by Megvii (Face++) and contains mostly images obtained from Google searches. Compared to other datasets, the CrowdHuman dataset exhibits a denser characteristic. The average number of objects per picture is significantly higher in this dataset compared to other datasets.

This dataset comprises 15,000 training pictures, 4,370 validation pictures, and 5,000 testing pictures. The training and validation images contained 470,000 instances, each picture contains an average of 23 individuals. However, the test pictures had no accompanying annotations. The Crowdhuman dataset is displayed in Figure 6.

The Caltech dataset is one of the datasets used for pedestrian detection tasks. This dataset is constitutive of 640x640 resolution, 30Hz videos captured in various scenes, including urban streets and campuses, etc. The training set comprises 42,782 images, and the test set comprises 4024 images. Figure 7 illustrates the Caltech dataset.

Figure 5
Partial data set display of Citypersons

Figure 6
Partial data set display of CrowdHuman

Figure 7
Partial data set display of Caltech

We use log-average miss rate based on false positives per image (FPPI) to calculate the proposed method's performance, which is determined by computing the geometric average of miss rates at 9 evenly spaced *FPPI* thresholds within the logarithmic range, specifically in the range of *FPPI* values from 0.01 to 1, the log-average miss rate determined by computing the mean is referred to as MR^{-2}. A lower value indicates superior model performance. When assessing model accuracy, the paper employs Precision (P), Recall (R), Average Precision (AP) index, AP@0.5(average accuracy at IoU=0.5), AP@0.5:0.95 (IoU between 0.5-0.95, step size 0.5), these equations are as shown below:

$$precision = \frac{TP}{TP + FP} \times 100\%, \tag{9}$$

$$recall = \frac{TP}{TP + FN} \times 100\%, \tag{10}$$

$$AP = \int_0^1 P(r)dr, \tag{11}$$

Where TP refers to True Positive, FP refers to False Positive, FN refers to False Negative.

4.2. Experimental Setup

The training method presented in this paper is as depicted: The initial learning rate is set to 10^{-2}, the learning rate is adjusted through cosine annealing decay, the optimization is performed using the SGD optimizer, and the batch size is configured as 32. The momentum is 0.937, and decay is configured as 0.0005. During the data preprocessing stage, random cropping is applied to the original images to obtain fixed-sized input images, followed by scaling and padding of the processed images.

4.3. Ablation Studies

In this section, to assess the influence of added components on the overall model and its effectiveness, we conducted ablation experiments utilizing the Citypersons datasets and evaluated its performance using MR^{-2}. The outcomes of the ablation test results for every component are displayed in Table 1.

To evaluate the impact of combining NWD with CIoU on pedestrian detection performance, and to find the optimal way of combining them, we tried for the first time to use NWD as a loss function instead of the traditional IoU. However, this change did not lead to an improvement in any performance. Therefore, we decided to continue using CIoU while fine-tuning it by adjusting the weight ratio between IoU loss and NWD. To achieve this objective, we designed six sets of comparison experiments and recorded the results in Table 2.

From Table 2, it is evident that different ratio relationships have meaningful effects on the detection performance of YOLOv7. The best detection results are achieved when the weight ratio of CIoU Loss to NWD is set to 0.7 and 0.3, respectively.

To verify the superiority of combining NWD and CIoU as the regression loss function, we employed several prevalent loss functions for comparison with our method on the YOLOv7 model. The experimental

Table 1

The contribution of every component, evaluated on the CityPersons dataset [50] under IoU=0.5, MR^{-2}lower is better

Method	DE-ELAN	Light-REFM	NWD CIoU	MR^{-2}(%)	Parameters(Mb)
YOLOv7				16.41	37.19
	✓			14.85	37.65
		✓		14.63	37.69
			✓	15.65	37.19
	✓	✓		14.17	38.15
		✓	✓	14.30	37.69
	✓		✓	14.45	37.65
	✓	✓	✓	13.83	38.15

Table 2
The outcomes of the various weight ratios between different CIoU Loss and NWD on the CityPersons dataset [50]

CIoU	NWD	AP50(%)	AP50:95(%)
1	0	61.93	36.18
0	1	50.62	25.54
0.5	0.5	64.31	39.18
0.6	0.4	65.48	39.75
0.4	0.6	58.32	33.82
0.7	0.3	66.54	41.15

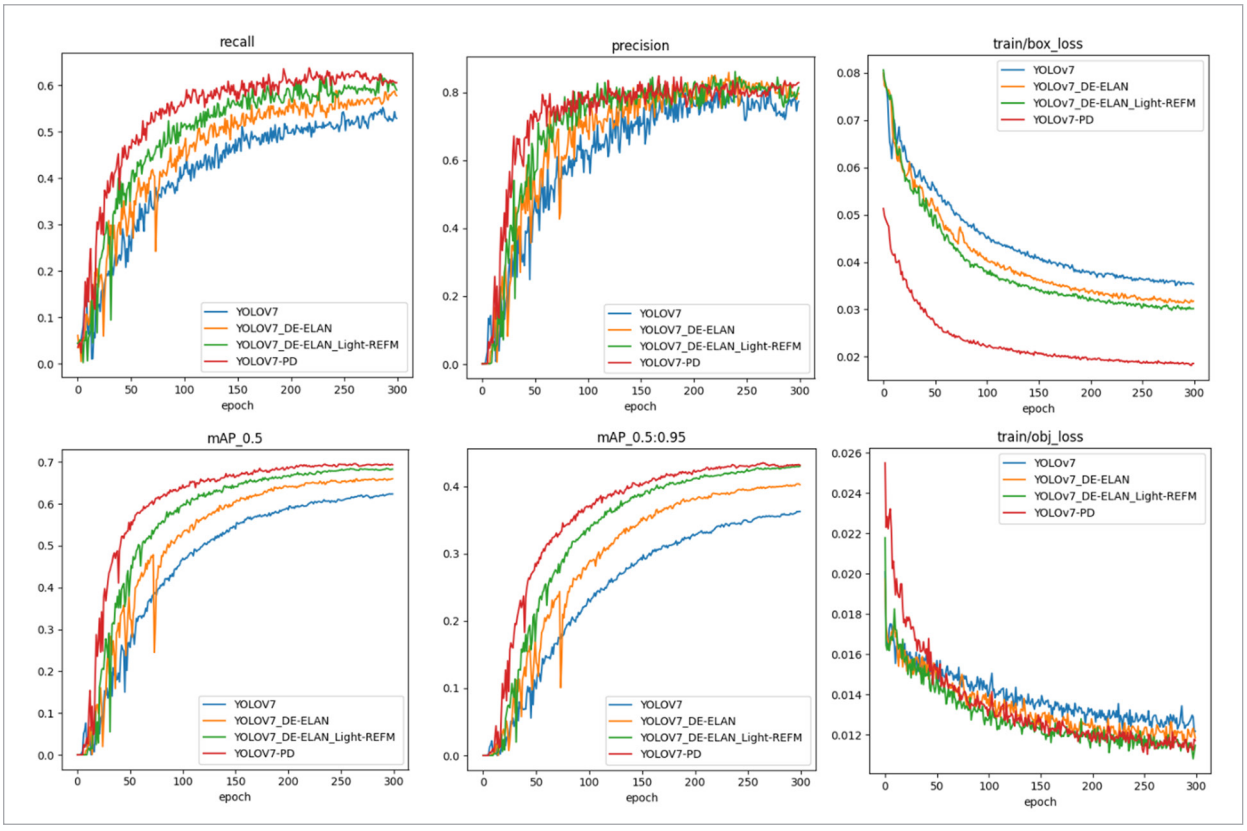
Table 3
Comparison of results using various loss functions on YOLOv7

Method	AP (%)	FPS
SIoU[17]	65.67	112
EIoU[51]	64.87	85
WIoU[43]	64.57	113
CIoU[57]	61.93	121
NWD-CIoU	66.54	116

results are presented in Table 3. These results clearly illustrate that integrating NWD and CIoU into the regression loss function can lead to a substantial enhancement in the model ability, effectively proving the validity of our method.

Furthermore, to achieve better compute the impact of addition components on the overall model accuracy. Figure 8 shows Precision, Recall, P-Rcurve, mAP@0.5, mAP@0.5:0.95, box loss, and object loss results obtained by our training on the Citypersons dataset. Compared to YOLOv7, YOLOv7-PD shows an

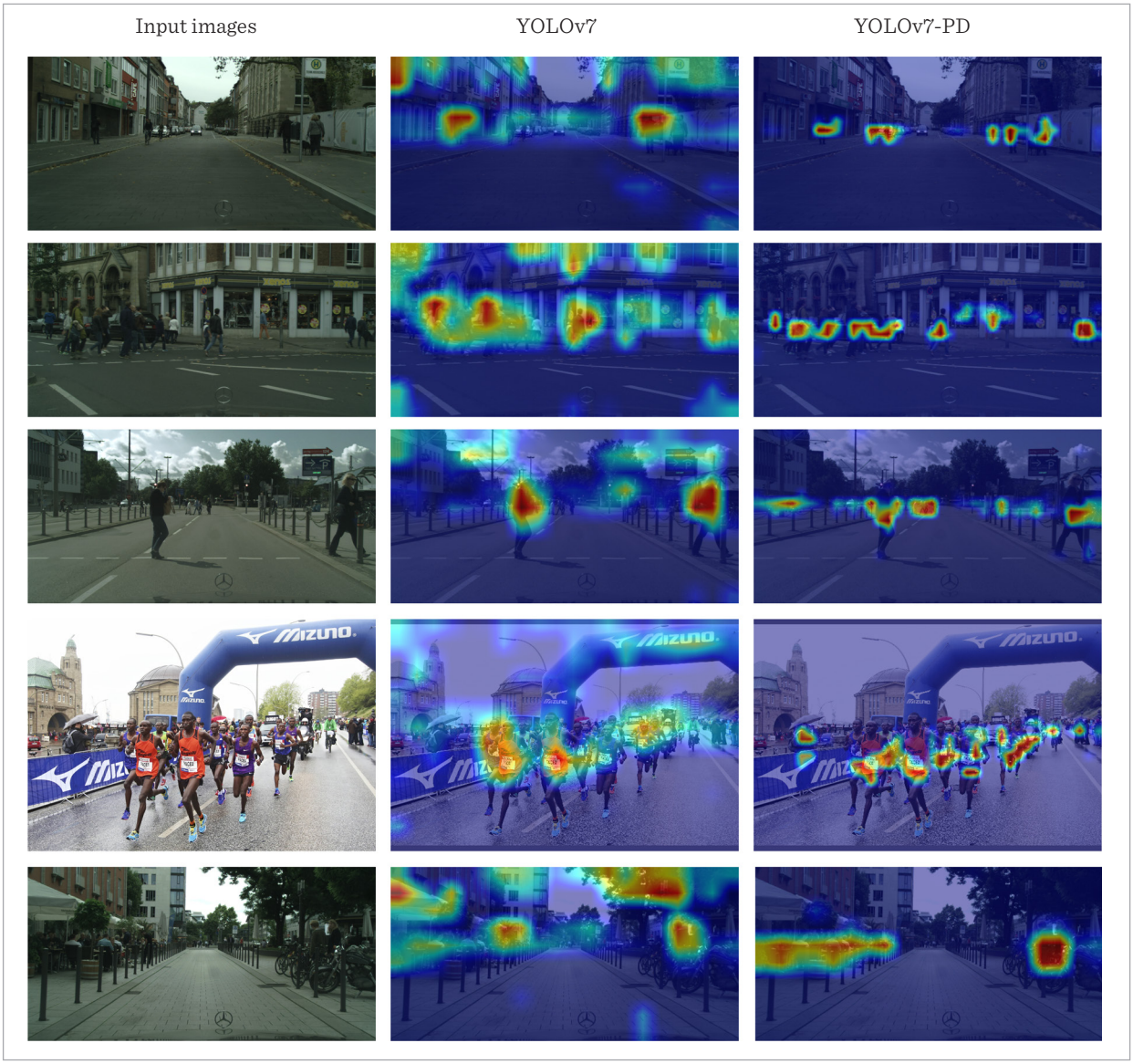
Figure 8
The results obtained by training on the Citypersons dataset



improvement of 5.51% in Precision, a 7.71% improvement in Recall, a 7.01% improvement in mAP@0.5, and a 6.88% improvement in mAP@0.5:0.95. YOLOv7-PD's box loss is consistently lower than the other three versions, indicating that YOLOv7-PD is more precise in locating the bounding boxes of objects. During training, YOLOv7-PD's object loss decreases relatively quickly and remains at a low. A lower box loss typically suggests that the model performs well in predicting the size and position of the targets. This indicates that YOLOv7-PD is more accurate in determining the presence of objects in images.

In Figure 9, we verify the effectiveness of YOLOv7-PD using Grad-CAM. The results are computed from the penultimate convolutional layer. To the left of every input image, the ground truth labels are displayed, and the distribution of the heatmap exhibits the distribution of interests within the network. We compared the

Figure 9
The visualization outcomes of Grad-CAM



visualizations of YOLOv7 and YOLOv7-PD, and these results were obtained by analyzing the penultimate convolutional layer. To the left side of every input picture, we displayed the ground truth labels, while the distribution of the heatmap reflected the spatial distribution of areas of interest within the network. The research findings indicate that YOLOV7-PD places more emphasis on the center of a pedestrian's torso. Furthermore, for smaller pedestrian targets, the focus area typically covers the entire body region. Compared to YOLOv7, our method exhibits superior performance in learning small pedestrian targets.

4.4. Comparison with the State-of-the-Art

In this section, we will assess the performance of the YOLOv7-PD model with some representative models on the Caltech, CityPersons, and CrowdHuman datasets. We will use consistent parameters, environment, and data processing techniques to compute the respective evaluation metrics.

For the purpose of validating the detection capabilities of YOLOv7-PD on pedestrian targets of varying scales, we chose several outstanding pedestrian detection models for comparison on subsets of the Caltech dataset separated into different scales. These include MSCM-ANet [34], which is designed with multi-scale convolution modules for extracting features at varying scales. RPN+BF [49] utilizes a Region Proposal Network to propose regions of interest, which are then processed by the backbone feature network to extract features relevant to subsequent object detection or classification tasks. TA-CNN [42] reduces the variance between datasets through a multi-task deep model. MHN [4] proposes a multi-branch and high-level semantic network that uses cross-layer connections to add context to a relatively smaller receptive field branch. It also incorporates and incorporates dilated convolutions to increase the output feature map's resolution. The Coupled Network [32] uses a gated multi-layer feature extraction subnetwork and a deformable region of interest pool to deal with occlusion issues in pedestrian detection. MS-CNN [3] proposes a multi-scale neural network for fast multi-scale target detection. SA-FastRCNN [25] solves the problem of multi-scale target detection by jointly training two networks for the detection of both large and small pedestrian targets. As displayed in Figure 10, when we tested on the reasonable subset of Caltech, YOLOv7-PD achieved an *MR* value of 15%, which is 2% lower than the second-best model. On the subsets with pedestrian heights of (30,80) and (80,inf), the *MR* values are 18% and 3% individually, which are at the forefront when compared to other methods. The *MR* value on the subset of pedestrian heights of (0, 30) is 21%, which is 1% lower than the sub-optimal model, demonstrating excellent performance. This achievement is mainly attributed to our design of the Light-REFM module, which finely captures multi-scale information by building a pyramid structure and employing dilation convolutions of different sizes. This structure enables the model to effectively recognize and process targets at different scales and is particularly good at detecting small targets and partial occlusion situations. This combination not only significantly improves the model's localization capability. In addition, we adopt a refined regression loss function derived from NWD and combine it with CIoU, an innovation that further enhances the model's performance concerning localization accuracy. But is also particularly effective for accurate detection and feature capture of small targets, thus ensuring efficient performance in complex scenes.

To evaluate the detection capability of YOLOv7-PD on pedestrian targets with different occlusion levels, we partitioned the CityPersons dataset into four subsets based on the level of occlusion. Reasonable subset: Pedestrians in this subset have a visibility range from 65% to 100%, encompassing individuals who are mostly visible with possible partial occlusions. Heavy Subset: The subset includes pedestrians with visibility less than 65%, representing significant occlusion or very challenging detection conditions. Partial subset: This subset is for pedestrians with visibility between 65% and 90%, indicating moderate occlusion. Bare subset: This subset includes pedestrians who are almost entirely visible, with visibility ranging between 95% and 100%. Using the same parameters, we compared YOLOv7-PD with the above methods. The results are presented in Table 4, In the bare subset, our method MR^{-2} achieves 6.7%. In the partial occlusion subset, our method MR^{-2} reaches 22.5%. In the heavy occlusion subset, our method MR^{-2} reaches 36.2%. The results indicate that YOLOv7-PD plays a significant role in detecting occluded pedestrian objects.

Figure 10

The *MR* and *FPPI* curves of the state-of-the-art method and the proposed YOLOv7-PD on the Caltech test set are presented in four different evaluation settings, a. Reasonable, b. Pedestrian height:(0,31), c. Pedestrian height:[31,81), d. Pedestrian height:(80,inf)

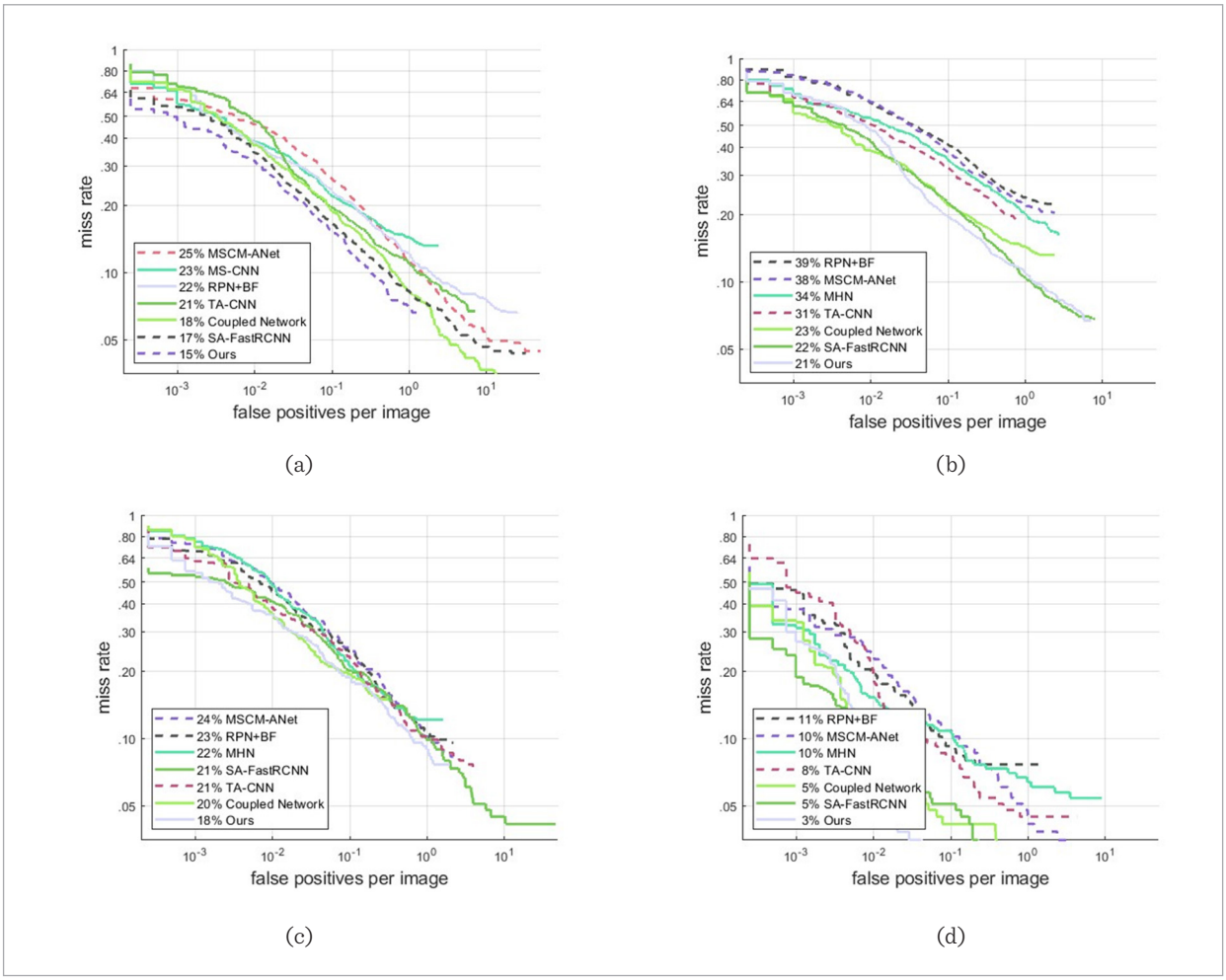


Table 4

Comparisons of YOLOv7-PD and other state-of-the-art methods on four Citypersons subsets

Method	Reasonable	Heavy	Partial	Bare
MSCM-ANet [34]	27.8	53.3	29.3	14.8
Adapt Faster-RCNN [7]	15.4	-	-	-
RPN+BF [49]	20.1	54.6	36.1	12.2
TA-CNN [42]	22.9	42.7	25.7	10.5
MHN [4]	19.4	42.4	25.4	9.2
Coupled Network [32]	16.1	36.4	23.8	7.8
Ours	15.2	36.2	22.5	6.7

Table 5
Comparison between YOLOv7-PD and other one-stage methods on Citypersons dataset

Method	Backbone	Size	FPS	AP(%)
RetinaNet [28]	ResNet50	640	43	52.23
RetinaNet [28]	ResNet101	640	34	54.17
YOLOv4 [1]	CSPDarknet53	640	72	59.04
YOLOX [16]	CSPDarknet53	640	78	61.52
Ours	CSPDarknet53(DE-ELAN)	640	96	69.31

In Table 5, we conducted AP values and FPS tests on YOLOv7-PD and some single-stage models to assess the models' effectiveness in pedestrian detection. In contrast to other single-stage methods, YOLOv7-PD showed a higher AP value, reaching 69.3%. Its FPS is 96. Our method shows significant advantages in different performance metrics, demonstrating that our method achieves excellent detection accuracy while sustaining a high processing speed.

To validate the detection capabilities of YOLOv7-PD for occluded objects, we opted for the CrowdHuman dataset, known for its abundance of occluded objects. We compared YOLOv7-PD against advanced methods, including RetinaNet [28], YOLOX [16], YOLO-CPD [15], GossipNet [20], RelationNet [21], MIP [8]. For assessment, we employed MR^{-2} and AP for evaluation, as depicted in Table 6. Our method AP achieves 83.36% and MR^{-2} achieves 41.2%, both of which reached the advanced level in the field of similar detection.

YOLOv7-PD achieves this result mainly due to the adoption of the improved DE-ELAN module, which utilizes ODConv and four complementary attentional mechanisms to effectively capture rich contextual information and enhance feature extraction. This enhanced feature extraction is essential for accurately detecting complex image details, especially in crowded or complex scenes. When there are extensive inter-class occlusions in the detection, the miss rate (MR^{-2}) becomes more crucial for evaluating the model.

In Figure 11, the detection outcomes of both YOLOv7 and YOLOv7-PD on the CityPersons dataset are displayed. We showcase extreme scenarios with small-scale objects and a high degree of occlusion. YOLOv7 exhibits challenges in detecting small pedestrian targets and those with a significant portion of their bodies occluded, resulting in missed detections and false alarms. In contrast, YOLOv7-PD demonstrates exceptional detection capabilities even when dealing with highly occluded pedestrians and small-sized pedestrians.

Table 6
Comparing various crowded detection methods using the validation set of CrowdHuman

Method	MR^{-2}(%)	AP(%)
RetinaNet [28]	63.33	80.8
YOLOX [16]	64.15	77.1
YOLO-CPD [15]	59.03	82.2
GossipNet [20]	49.4	80.4
RelationNet [21]	48.2	81.6
MIP [8]	42.8	86.7
Ours	41.2	83.36

5. Conclusion

In this paper, we propose three improvement strategies aimed at enhancing the performance of YOLOv7 in pedestrian detection. First, we propose the ODConv based module DE-ELAN, which considers dynamics in spatial dimensions, input channels, and output channels. This module enhances the model's capability to extract features, particularly for small-scale and occluded pedestrians, by effectively suppressing interference noise associated with background features and strengthening critical feature information.

Figure 11
The detection results of YOLOv7 and YOLOv7-PD using the CityPersons dataset

(a) Input images (b) YOLOv7 (c) YOLOv7-PD

Second, we propose Light-REFM, which groups input mappings in channel sequences and extracts features through dilated convolutions to obtain richer contextual information. Finally, we combine the traditional CIoU method with NWD to enhance the YOLOv7's performance in small object and large object detection tasks, enabling the model to better adapt to different scales of objects. The effectiveness of this method has been validated by comparing it to other excellent methods on three various datasets. The research results indicate that our approach demonstrates a certain level of competitiveness concerning accuracy and robustness.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (62262004) (http://www.nsfc.gov.cn). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Bochkovskiy, A., Wang, C. Y., Liao, H. Y. M. Yolov4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934, 2020. https://doi.org/10.48550/arXiv.2004.10934
2. Brauwers, G., Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4), 3279-3298. https://doi.org/10.1109/TKDE.2021.3126456
3. Cai, Z., Fan, Q., Feris, R. S., Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016: 354-370. https://doi.org/10.1007/978-3-319-46493-0_22
4. Cao, J., Pang, Y., Zhao, S., Li, X. High-level Semantic Networks for Multi-scale Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(10), 3372-3386. https://doi.org/10.1109/TCSVT.2019.2950526
5. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184
6. Chen, R., Wang, X., Liu, Y., Wang, S., Huang, S. A Survey of Pedestrian Detection Based on Deep Learning. Communications, Signal Processing, and Systems: Proceedings of the 8th International Conference on Communications, Signal Processing, and Systems 8th. Springer Singapore, 2020, 1511-1516. https://doi.org/10.1007/978-981-13-9409-6_181
7. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 3339-3348. https://doi.org/10.1109/CVPR.2018.00352
8. Chu, X., Zheng, A., Zhang, X., Sun, J. Detection in Crowded Scenes: One Proposal, Multiple Predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12214-1222. https://doi.org/10.1109/CVPR42600.2020.01223
9. Combs, T. S., Sandt, L. S., Clamann, M. P., McDonald, N. C. Automated Vehicles and Pedestrian Safety: Exploring the Promise and Limits of Pedestrian Detection. American Journal of Preventive Medicine, 2019, 56(1), 1-7. https://doi.org/10.1016/j.amepre.2018.06.024
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 3213-3223. https://doi.org/10.1109/CVPR.2016.350
11. Dalal, N., Triggs, B. Histograms of Oriented Gradients for Human Detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1, 886-893. https://doi.org/10.1109/CVPR.2005.177
12. Devipriya, A., Prabakar, D., Singh, L., Oliver, A. S., Qamar, S., Azeem, A. Machine Learning-Driven Pedestrian Detection and Classification for Electric Vehicles: Integrating Bayesian Component Network Analysis and Reinforcement Region-based Convolutional Neural Networks. Signal, Image and Video Processing, 2023, 17(8), 4475-4483. https://doi.org/10.1007/s11760-023-02681-1
13. Dollár, P., Wojek, C., Schiele, B., Perona, P. Pedestrian Detection: A Benchmark. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, 304-311. https://doi.org/10.1109/CVPRW.2009.5206631
14. Fang, F., Liang, W., Cheng, Y., Xu, Q., Lim, J. H. Enhancing Representation Learning with Spatial Transformation and Early Convolution for Reinforcement Learning-based Small Object Detection. IEEE Transactions on Circuits and Systems for Video Technology, 2023. https://doi.org/10.1109/TCSVT.2023.3284453
15. Gao, F., Cai, C., Jia, R., Hu, X. Improved YOLOX for Pedestrian Detection in Crowded Scenes. Journal of Real-Time Image Processing, 2023, 20(2), 24. https://doi.org/10.1007/s11554-023-01287-7
16. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. Yolox: Exceeding Yolo Series in 2021. arXiv preprint arXiv: 2107.08430, 2021. https://doi.org/10.48550/arXiv.2107.08430
17. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. arXiv preprint arXiv:2205.12740, 2022. https://doi.org/10.48550/arXiv.2205.12740
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision. 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169
19. He, K., Zhang, X., Ren, S., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

20. Hosang, J., Benenson, R., Schiele, B. Learning Non-maximum Suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 4507-4515. https://doi.org/10.1109/CVPR.2017.685

21. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y Relation Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 3588-3597. https://doi.org/10.1109/CVPR.2018.00378

22. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

23. Khan, A. H., Nawaz, M. S., Dengel, A. Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023, 5476-5485. https://doi.org/10.1109/CVPR52729.2023.00530

24. Li, C., Zhou, A., Yao, A. Omni-Dimensional Dynamic Convolution. arXiv preprint arXiv:2209.07947, 2022. https://doi.org/10.48550/ arXiv:2209.07947

25. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S. Scale-Aware Fast R-CNN for Pedestrian Detection. IEEE Transactions on Multimedia, 2017, 20(4), 985-996. https://doi.org/10.1109/TMM.2017.2759508

26. Li, R., Zu, Y. Research on Pedestrian Detection Based on the Multi-Scale and Feature-Enhancement Model. Information, 2023, 14(2), 123. https://doi.org/10.3390/info14020123

27. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

28. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

29. Liu, M., Jiang, J., Zhu, C., Yin, X. C. VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 6662-6671. https://doi.org/10.1109/CVPR52729.2023.00644

30. Liu, M., Wan, L., Wang, B., Wang, T. SE-YOLOv4: Shuffle Expansion YOLOv4 for Pedestrian Detection Based on PixelShuffle. Applied Intelligence, 2023, 53(15), 18171-18188. https://doi.org/10.1007/s10489-023-04456-0

31. Liu, S., Huang, D. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), 2018, 385-400. https://doi.org/10.1007/978-3-030-01252-6_24

32. Liu, T., Luo, W., Ma, L., Huang, J. J., Stathaki, T., Dai, T. Coupled Network for Robust Pedestrian Detection with Gated Multi-Layer Feature Extraction and Deformable Occlusion Handling. IEEE Transactions on Image Processing, 2020, 30, 754-766. https://doi.org/10.1109/TIP.2020.3038371

33. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. SSD: Single Shot Multibox Detector. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

34. Ma, J., Wan, H., Wang, J., Xia, H., Bai, C. An Improved One-Stage Pedestrian Detection Method Based on Multi-Scale Attention Feature Extraction. Journal of Real-Time Image Processing, 2021, 1-14. https://doi.org/10.1007/s11554-021-01074-2

35. Misra, D., Nalamada, T., Arasanipalai, A. U., Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021, 3139-3148. https://doi.org/10.1109/WACV48630.2021.00318

36. Mo, W., Zhang, W., Wei, H., Cao, R., Ke, Y., Luo, Y. PV-Det: Towards Pedestrian and Vehicle Detection on Gigapixel-Level Images. Engineering Applications of Artificial Intelligence, 2023, 118, 105705. https://doi.org/10.1016/j.engappai.2022.105705

37. Niu, Z., Zhong, G., Yu, H. A Review on the Attention Mechanism of Deep Learning. Neurocomputing, 2021, 452, 48-62. https://doi.org/10.1016/j.neucom.2021.03.091

38. Pan, M., Chen, J., Wang, S., Dong, Z. A Novel Approach for Marine Small Target Detection Based on Deep Learning. 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP). IEEE, 2019, 395-399. https://doi.org/10.1109/SIPROCESS.2019.8868862

39. Ren, J., Ren, R., Green, M., Huang, X. Defect Detection from X-Ray Images Using A Three-Stage Deep Learning Algorithm. In 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE). IEEE, 2019, 1-4. https://doi.org/10.1109/CCECE.2019.8861944

40. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J. Crowdhuman: A Benchmark for Detecting Human in a

Crowd. arXiv preprint arXiv:1805.00123, 2018. https://doi.org/10.48550/arXiv.1805.00123

41. Song, X., Chen, B., Li, P., He, J. Y., Wang, B., Geng, Y., Zhang, H. Optimal Proposal Learning for Deployable End-to-End Pedestrian Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 3250-3260. https://doi.org/10.1109/CVPR52729.2023.00317
42. Tian, Y., Luo, P., Wang, X., Tang, X. Pedestrian Detection Aided by Deep Learning Semantic Tasks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 5079-5087. https://doi.org/10.1109/CVPR.2015.7299143
43. Tong, Z., Chen, Y., Xu, Z., Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. arXiv preprint arXiv:2301.10051, 2023. https://doi.org/10.48550/ arXiv: 2301.10051
44. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721
45. Wang, J., Xu, C., Yang, W., Yu, L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. arXiv preprint arXiv:2110.13389, 2021. https://doi.org/10.48550/arXiv.2110.13389
46. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155
47. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
48. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D. EPSANET: An Efficient Pyramid Split Attention Block on Convolutional Neural Network. arXiv preprint arXiv:2105.14447, 20212. https://doi.org/10.1007/978-3-031-26313-2_33
49. Zhang, L., Lin, L., Liang, X., He, K. Is Faster R-CNN Doing Well for Pedestrian Detection?. In Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer International Publishing, 2016, 443-457. https://doi.org/10.1007/978-3-319-46475-6_28
50. Zhang, S., Benenson, R., Schiele, B. CityPersons: A Diverse Dataset for Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 3213-3221. https://doi.org/10.1109/CVPR.2017.474
51. Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T. Focal and Efficient IOU Loss For Accurate Bounding Box Regression. Neurocomputing, 2022, 506, 146-157. https://doi.org/10.1016/j.neucom.2022.07.042
52. Zheng, Q., Tian, X., Yu, Z., Jiang, N., Elhanashi, A., Saponara, S., Yu, R. Application of Wavelet-Packet Transform Driven Deep Learning Method in PM2. 5 Concentration Prediction: A Case Study of Qingdao, China. Sustainable Cities and Society, 2023, 92, 104486. https://doi.org/10.1016/j.scs.2023.104486
53. Zheng, Q., Tian, X., Yu, Z., Wang, H., Elhanashi, A., Saponara, S. DL-PR: Generalized Automatic Modulation Classification Method Based on Deep Learning with Priori Regularization. Engineering Applications of Artificial Intelligence, 2023, 122, 106082. https://doi.org/10.1016/j.engappai.2023.106082
54. Zheng, Q., Zhao, P., Li, Y., Wang, H., Yang, Y. Spectrum Interference-Based Two-Level Data Augmentation Method in Deep Learning for Automatic Modulation Classification. Neural Computing and Applications, 2017, 33(13), 7723-7745. https://doi.org/10.1007/s00521-020-05514-1
55. Zheng, Q., Zhao, P., Wang, H., Elhanashi, A., Saponara, S. Fine-Grained Modulation Classification Using Multi-Scale Radio Transformer with Dual-Channel Representation. IEEE Communications Letters, 2022, 26(6), 1298-1302. https://doi.org/10.1109/LCOMM.2022.3145647
56. Zheng, Q., Zhao, P., Zhang, D., Wang, H. MR-DCAE: Manifold Regularization-Based Deep Convolutional Autoencoder for Unauthorized Broadcasting Identification. International Journal of Intelligent Systems, 2021, 36(12), 7204-7238. https://doi.org/10.1002/int.22586
57. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07), 12993-13000. https://doi.org/10.1016/j.cose.2022.102748