IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 June 2023, accepted 3 July 2023, date of publication 6 July 2023, date of current version 20 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292881

RESEARCH ARTICLE

PEFNet: Position Enhancement Faster Network for Object Detection in Roadside Perception System

LEI HUANG[1], WENZHUN HUANG[1], HAI GONG[1], CHANGQING YU[1], AND ZHUHONG YOU[2]

[1]School of Electronic Information, Xijing University, Xi'an, Shaanxi 710123, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi 710123, China

Corresponding author: Wenzhun Huang (huangwenzhun@xijing.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 62072378.

ABSTRACT Roadside perception is a challenging research area that presents even greater difficulties than vehicle perception. Due to the different locations and angles of cameras, roadside objects exhibit violent multi-scale variations, while the vast sensing field introduces more small-scale targets and complex backgrounds, making target recognition more challenging. To address these problems, we focus on position information encoding to achieve accurate roadside object detection by proposing the position enhancement faster network (PEFNet). Based on YOLOv6, the FasterNet Block is introduced into Backbone and Neck networks to provide efficient feature extraction while achieving model lightweight transformation. To improve small target detection performance, a position-aware feature pyramid network (PA-PAN) is proposed to enhance position information encoding, and the SPD-Conv is applied in the PA-PAN to further enhance effective feature extraction. Finally, the TSCODE is integrated into the detection head to achieve accurate target recognition and suppress background noise interference. Experiments on the Rope3D and UA-DETRAC datasets show that our model outperforms advanced YOLOv6, YOLOX, and FCOS in roadside object detection. Compared with YOLOv6, our method improves the mAP0.50 on the Rope3D dataset from 78.18% to 82.39%, with the AP of small objects such as pedestrians increasing by 7.01%. Furthermore, PEFNet reduces the weight of the network by 43.1% while maintaining detection speed at 75fps and achieving higher accuracy than previous algorithms for the same number of frames.

INDEX TERMS Feature extraction, position enhancement, feature aggregation, decoupled head, roadside images, object detection.

I. INTRODUCTION

Roadside perception is an essential technology for intelligent transportation systems to achieve vehicle-road collaborative perception. Recent traffic accidents have highlighted the limitations of vehicle-side perception algorithms, which have restricted sensing range, and are subject to the influence of obstacles, adverse weather, lighting, and the surface reflectivity and motion state of the objects being sensed [1]. To improve driving safety, roadside perception can provide real-time information about vehicles and pedestrians [2]. Furthermore, with the development of edge intelligence, roadside perception can further enhance safety by fusing lidar, camera, and ultrasonic sensing technologies, and providing accurate and real-time information, while processing data in real-time to improve the response speed of roadside perception and better ensure driving safety [3], [4].

The main purpose of roadside perception is to improve the beyond-line-of-sight perception capability of intelligent vehicles, extend the perception range of intelligent networked vehicles, and provide early warning [5], [6], [7]. Currently, roadside perception technologies use visual, auditory, and other sensors, as well as recognition and behavior recognition technologies, to detect the environment and traffic

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

participants around the road. This technology can help to accurately measure the position, speed, and direction of moving targets (such as pedestrians, vehicles, etc.) and static obstacles on the road to realize the automatic driving of vehicles and the coordination of road traffic. Therefore, roadside perception technologies not only improve the efficiency of road traffic but also ensure the safety of drivers and passengers.

Computer vision-based object detection algorithms are essential for roadside perception, as they are tasked with accurately locating and identifying targets in collected images. However, compared with the object detection task based on vehicle-side views, roadside object detection tends to meet more challenges. First, differences in the location and height of the roadside camera installation and the camera shooting angle lead to violent object scale variance. Second, roadside detection algorithms typically cover a larger area, resulting in more detection targets and complex backgrounds in the captured images. Third, the complex road environment and the interference of natural factors such as lighting bring great challenges to the generalizability of the object detection model. Finally, the efficiency of the detection model is of paramount importance, as it must quickly and accurately feed information to the intelligent vehicle. The above problems make it impossible to apply the vehicle-side object detection algorithm directly to roadside object detection, so it is crucial to design an object detection model with real-time detection and high generalization capability for roadside object detection [8].

In recent years, object detection has been one of the most active and challenging research fields in computer vision due to its wide range of applications and complex scenarios. Therefore, researchers have proposed and developed various advanced object detection methods [9], [10], [11]. Particularly for industrial sectors that require high-performance object detection methods with real-time constraints, one-stage detectors [5], [12], [13] with efficient network architectures [14], [15], [16] and advanced training stages [17], [18], [19] such as YOLOv3 [20], YOLOX [21] and YOLOv7 [22] have been developed. These methods achieve better AP-Latency balance on the COCO dataset [23]. Network architecture plays a vital role in object detection. Historically, Darknet has been dominant [24], but other effective detection networks have been studied, such as YOLOv6 [25] and DAMO-YOLO [26]. The YOLOv6 utilizes RepVGG [27] to design a hardware-aware network architecture EfficientRep, which effectively utilizes hardware computing power and memory bandwidth. Feature Pyramid Network (FPN) [28] has been demonstrated to be an effective way to fuse multi-scale features, and recently Jiang et al. [29] proposed a Generalized-FPN (GFPN), which further improves the FPN performance through a novel queen-fusion method, but at the cost of a large amount of computation. The ERepGFPN is proposed in DAMO-YOLO for further decoupling and optimizing GFPN. These advanced target detection methods have achieved better results in various scenarios.

However, for roadside scenes with dramatic object scale variations and dense small targets, the existing networks pay less attention to fine-grained features and are prone to lose the feature information of small targets during the network model training iterations. Moreover, in the dense target environment, the overlap and coverage of features pose a great challenge to object distinction. We found that by aggregating shallow-level detail features across scales, the network's attention to fine-grained features can be enhanced. Therefore, developing a method to explore the correlation between multi-scale feature maps for better acquisition and aggregation of effective feature information is of great significance in improving the performance of roadside object detection algorithms.

As aforementioned, current object detection algorithms are designed for natural scenes. However, in the roadside object detection task, more challenging tasks such as violent object scale variance, small objects, and dense objects make the current algorithms not directly usable. Furthermore, the increasing complexity of the model network makes it difficult to deploy them on roadside edge devices. To address the limitations of existing models, this paper investigates the advanced detector YOLOv6, which combines the strengths of YOLOv5 and YOLOX and outperforms other algorithms of similar size in terms of both accuracy and speed. Based on the characteristics of roadside object detection, this paper optimizes YOLOv6 and proposes the position enhancement faster network (PEFNet).

Our contributions are listed as follows:

1) To improve small target detection accuracy, the position-aware feature pyramid network (PA-PAN) is proposed to obtain more effective small-scale features contained in roadside images by enhancing the position information encoding.
2) The PEFNet structure combines FasterNet Block and SPD-Conv, which can effectively capture roadside image fine-grained features and enhance the position correlation between features, and compress the model volume at the same time.
3) For poor detection performance in complex roadside background noise, the TSCODE is integrated into the detection head to achieve accurate target recognition and suppress background noise interference, by decoupling localization and classification tasks.
4) On the Rope3D and UA-DETRAC test datasets, the proposed PEFNet and advanced detectors are evaluated. Compared with current state-of-the-art detectors, the proposed PEFNet shows great potential in terms of model lightweight and detection performance.

This letter is organized as follows: In Section II, related works on efficient convolutional strategies, multi-scale feature aggregation, and object detection models are introduced. In Section III, the novel roadside object detection method is proposed. In Section IV, details of experiments and comparison results are provided. Finally, conclusions and suggestions are given in Section V.

II. RELATED WORKS

Traditional YOLO models are designed for object detection tasks in natural scenes, and directly using these models for object detection on roadside images has several major issues [30], as shown in some cases in Figure 1. Firstly, due to the difference in the position and height of the roadside cameras and the angle of camera shooting, the scale of objects varies violently, which may lead to false detection and omission. Secondly, the captured images often contain more low-resolution and blurred targets due to interference from lighting and natural factors; especially under complex traffic conditions, the overlapping and occlusion of dense objects further increase the difficulty of detection. Thirdly, due to the larger coverage of the roadside view, the images obtained from roadside sensors contain complex backgrounds and a large number of extremely small-sized objects, which are not easily recognizable. These issues lead to poor performance of traditional YOLO models in roadside scene images and thus cannot be directly applied to roadside scene object detection tasks.

FIGURE 1. Examples are given to illustrate the three main problems of object detection on roadside images. The cases in the first, second, and third rows show the problems of object size variation, tiny object sizes, and natural factors interference that are difficult to detect, respectively.

With the development of deep neural networks, more and more optimization strategies are proposed to address problems in specific environments. To improve the performance of object detection and address the issues in roadside object detection, we analyzed and studied the latest progress in deep learning [24], [31]. We found that the current main detection algorithms for improvement are based on convolutional neural networks (CNN) [32], [33], [34]. Despite its effectiveness in extracting feature information, the limitation of the convolution structure restricts its ability to obtain global context information. To this end, many strategies to improve convolution have been proposed [35], [36], [37], [38]. In addition, multi-scale feature fusion, as an effective method to improve network performance, can improve the extraction of effective features, but its complex structure tends to introduce more computation, which brings challenges to edge devices with limited computing resources. Therefore, we discuss and analyze the current advanced convolutional strategies, multi-scale feature fusion, and object detection models to design a more efficient and accurate network model for roadside object detection.

A. EFFICIENT CONVOLUTIONAL STRATEGIES

Convolutional neural networks (CNN) have achieved great success in computer vision tasks such as image classification and object detection. However, their performance drops rapidly in more challenging tasks such as low-resolution images and small objects. This is due to the inherent defects of CNNs in feature learning, which originate from the inherent geometry of the CNN module: the convolutional unit samples the feature map at a fixed position; the pooling layer pools with a fixed ratio. To this end, many advanced convolutional strategies have been proposed. For instance, to tackle the problem of losing fine-grained information in the convolution process, Sunkara and Luo [35] proposed SPD-Conv to improve the extraction of effective features by replacing downsampling with space-to-depth (SPD) layers while preserving all channel information. It was applied to both YOLOv5 and ResNet and showed impressive performance. In addition, for objects with complex geometric variations, Dai et al. [36] proposed a DCN module that shifts the sampling points of the feature map by introducing an offset to increase attention to the important information. Recently, DCNv2 [37] was proposed to further optimize the efficiency of key region feature extraction. Experiments on the VOC [39] and COCO [23] datasets show that the models improved by DCN have achieved better performance than the original models. In conclusion, considering the uniqueness of roadside images, introducing efficient convolutional strategies is an effective way to improve the performance of roadside object detection.

B. MULTISCALE FEATURE AGGREGATION

Recently, several works have been proposed to improve the performance of small object detection for roadside perception [5], [40], [41]. Among them, multi-scale feature aggregation (MSFA) is a popular approach. MSFA effectively improves the performance of small object detection by exploiting multi-scale features from multiple layers of a deep convolutional neural network (DCNN). In MSFA, the features from each layer are combined with the features from the other layers by using a specific fusion operation, such as summation or con-catenation. The fused features are then used to refine the detection results. For example, Deng et al. [41] proposed an Extended Feature Pyramid Network (EFPN) approach to combine the features from different layers of a DCNN to improve the detection performance on small objects. The authors applied their method to the

Tsinghua-Tencent 100 K object detection dataset and achieved a better performance than previous methods. Wu et al. [40] proposed a multi-scale feature aggregation module to address the issue of scale variation and established a cross-scale refinement module to obtain more effective multi-scale features. Experiments demonstrate that the model outperforms the latest state-of-the-art detectors on five benchmark datasets. Chu et al. [42] proposed a multi-layer convolution feature fusion (MCFF) to improve multi-scale object detection performance, by fusing high-level and low-level features. The testing results on the Kitti dataset showed that the improved model exhibited better performance and generalization ability. In conclusion, multi-scale feature aggregation is an effective approach to improve the performance of small object detection for roadside perception. By combining the features from multiple layers of a DCNN, this approach can effectively exploit the multi-scale features to refine the detection results.

C. OBJECT DETECTION MODELS

Deep learning-based object detection has become a hot spot in the field of computer vision, playing an important role in object recognition and tracking. With the deepening of research, object detection based on deep learning can be mainly divided into two basic paradigms: anchor-based detection methods and an-chor-free detection methods. Anchor-based detection methods, such as Faster R-CNN [43] and YOLO [44], [20], [45] use predefined anchor boxes to compare with the object positions in the input image, which are defined by their center coordinates, width, height, and aspect ratio. Then this model recognizes the objects and their corresponding bounding boxes by optimizing the joint objective function, which measures the overlap between the ground truth and the predicted boxes. These models have achieved impressive results on various datasets, such as VOC and COCO. Anchor-free detection methods, such as FCOS [12] and CenterNet [11], do not use predefined anchor boxes. Instead, they directly predict the bounding boxes without needing massive predefined anchor boxes, making the model complexity lower and the detection performance more stable, which is currently a cutting-edge approach in object detection. In recent years, in order to further improve the object detection performance in general scenes, novel one-stage object detection networks such as YOLOX [21] and YOLOv6 [25] have been proposed. They adopted the advanced anchor-free paradigm and optimized the network structure and performance. However, due to the use of large-capacity feature extraction backbone networks and feature extraction modules, the computational cost has significantly increased, making it difficult to directly apply to real-time detection tasks in roadside scenes with limited hardware resources. At the same time, the performance of YOLOv6 and YOLOX for dense target and multi-type roadside environment detection still needs to be verified.

III. METHODOLOGY

In this section, we present a detailed description of the components of PEFNet, including the FasterNet Backbone, the PA-PAN Neck, and the TSCODE Head. The whole framework of PEFNet is illustrated in Figure 2.

A. PEFNET NETWORK ARCHITECTURE

To address the problem of limited feature extraction, complex backgrounds, and small target, which makes it difficult to simultaneously improve the detection speed and accuracy of roadside objects, this paper proposes a position enhancement faster network (PEFNet) based on YOLOv6 for roadside object detection. The network architecture of the PEFNet is shown in Figure 2.

According to the characteristics of roadside images, in PEFNet, we integrate the current advanced improvement strategies to design a more efficient and generalized network. PEFNet consists of three parts: the Backbone network for feature extraction, the Neck network for feature fusion, and the Head network for detection result generation. Firstly, FasterNet Block is introduced into the Backbone and Neck networks to provide efficient feature extraction and achieve model lightweight transformation. Subsequently, different scale feature maps extracted from the Backbone network are fed into the Neck network for feature fusion. Since previous works have not paid enough attention to fine-grained features, we propose a novel PA-PAN to replace PANet for feature aggregation, which improves the detection accuracy of small targets. Specifically, based on the original top-down and bottom-up information transmission paths, and to make the model focus more on position information encoding, the PEFA is introduced into PA-PAN to add an information transmission path. The PEFA guides shallow features with rich position information to flow back into the next layer feature map, and it can compensate for the loss of detailed information during the convolution process, improving the multi-scale feature generalization ability. Furthermore, the SPD module is introduced to retain more effective feature information, and further enhance the model's position information encoding. Finally, the output of Neck network is fed into the Head network for classification and localization prediction, and to improve object recognition and localization accuracy, TSCODE is added to the Head network. The TSCODE is used to feed specific features to the respective task branch, and it can further decouple the classification and localization tasks, maximizing the performance of the decoupled head.

B. POSITION-AWARE FEATURE PYRAMID NETWORK

Multiscale feature aggregation has been shown to be an effective component for object detection, and the representative algorithms are the feature pyramid network (FPN) [28], path aggregation network (PANet) [46], and bidirectional feature pyramid network (BiFPN) [15]. Their excellent performance

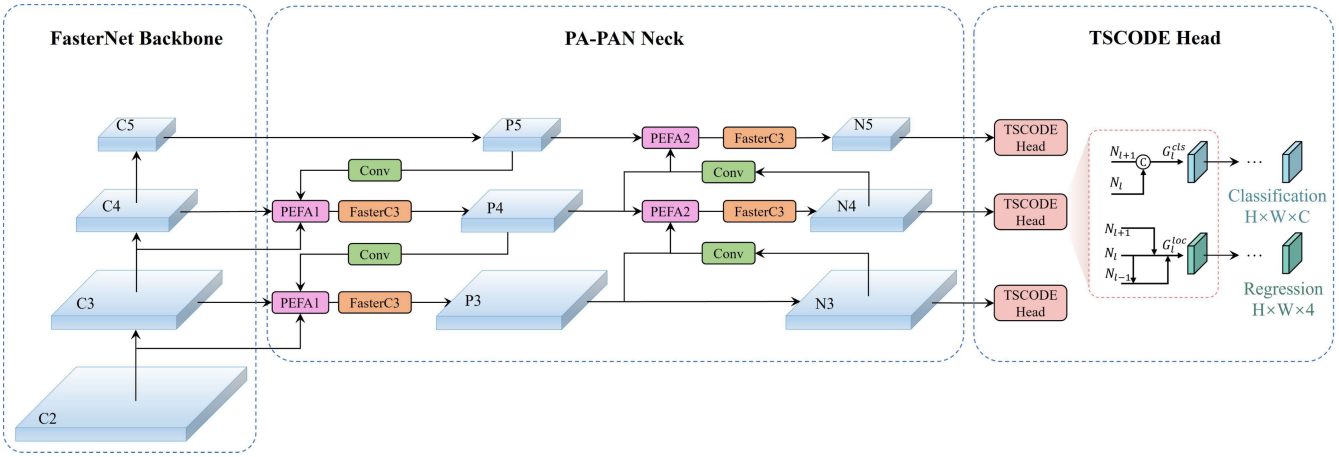


FIGURE 2. Illustration of the proposed PEFNet network architecture. 1) The FasterNet Block is introduced into Backbone and Neck networks for the model lightweight transformation. 2) The PANet is replaced by PA-PAN to integrate semantic and detailed information, including enhanced position information. 3) The TSCODE is inserted into the Head to further decouple the classification and localization tasks. 4) The SPD-Conv is added to the network to enhance effective feature extraction.

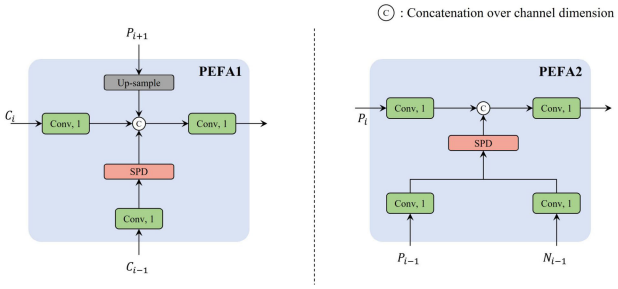


FIGURE 3. The feature aggregation module focuses on the position information of shallow features. The PEFA1 module is used for feature aggregation in the upsampling stage. The PEFA2 is used for feature aggregation in the downsampling stage.

in object detection and target segmentation tasks has left a deep impression on people [5], [13]. In recent years, many novel FPN structures have been proposed to adapt to complex scenarios in object detection tasks, such as GFPN [29] and ERepGFPN [26].

Although the performance of advanced FPN structures on many object detection tasks is impressive, they inevitably introduce more computation and parameters to aggregate more feature information, leading to a greatly increased complexity of the model. Furthermore, the performance of current FPN structures has significantly decreased in scenarios of small targets and complex backgrounds. This is mainly due to the fact that the current network pays less attention to fine-grained features, causing a large number of small target features to be lost in the convolution process and aggregating a large amount of redundant information [35]. Therefore, it is of great significance to design an efficient and lightweight FPN structure for roadside object detection.

Motivated by the fact that shallow features contain rich position and detail information, it is intuitive to aggregate shallow features to compensate for the loss of small target features, and thus improve the network's attention to fine-grained targets. Based on the previous work, we designed a position-aware feature pyramid network (PA-PAN) as the Neck of our network, as shown in Figure 2. Compared to the original structure, adding an extra path for aggregating shallow features. In Figure 3, the position enhancement feature aggregation (PEFA) module, as the core of the PA-PAN structure, can aggregate the feature maps of three adjacent layers at the same time. The PEFA module consists of PEFA1 in the upsampling stage and PEFA2 in the downsampling stage. It is worth mentioning that we also use the SPD module to downsample the shallow features while preserving all channel information, thus aggregating more effective details.

The PEFA1 is calculated as follows:

$$P_i = \text{Concat}(\text{Conv}(C_i), \mu(P_{i+1}), \varphi_{spd}(\text{Conv}(C_{i-1}))) \quad (1)$$

where Concat$(\cdot)$ represents channel concatenation operation; Conv$(\cdot)$ represents a convolutional layer; $\mu(\cdot)$ represents upsampling; $\varphi_{spd}(\cdot)$ represents SPD module, and it can be used for downsampling while preserving all channel features. The output P_i is used in PEA2 of the next stage, and PEFA2 is calculated as follows:

$$F_i = \text{Concat}(\text{Conv}(P_i), \varphi_{spd}(\text{Conv}(P_{i-1}, N_{i-1}))) \quad (2)$$

where, P_{i-1} and N_{i-1} are aggregated into the network after feature enhancement by the SPD module, thus enhancing the network position information encoding.

C. SPD-CONV MODULE

Compared to the object detection task based on the vehicular view, the object detection task based on the roadside view has more challenges. This is because the roadside view has a larger perception field, resulting in images acquired through roadside sensors containing more as well as complex information. Therefore, the object detection model needs to face more challenges, such as dense targets, target occlusion, and

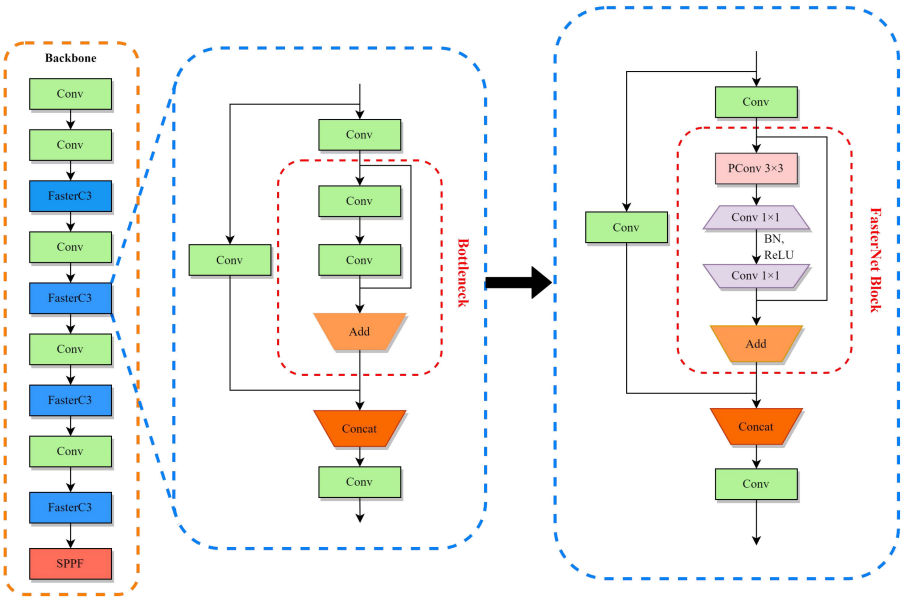


FIGURE 4. The improved Backbone structure by introducing FasterNet Block. The Bottleneck of the C3 module is replaced with FasterNet Block.

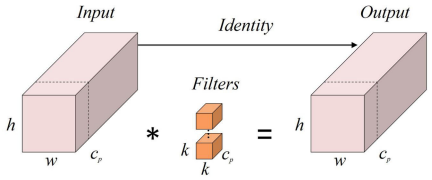


FIGURE 5. The overall architecture of PConv. Applying filters on only a few input channels effectively reduces redundant calculations and improves implementation efficiency.

natural scale variation. In addition, due to the interference of natural factors, the complex and changeable road environment has also brought great difficulties to object detection.

Although the convolutional neural network models have performed impressively in many computer vision tasks, their performance quickly declines in image resolution low or small target detection tasks. In short, this is a weakness of the CNN structures, that using convolution for feature extraction inevitably leads to the loss of fine-grained information and low feature representation learning efficiency [36]. Therefore, how to prevent or reduce the loss of small target feature information in the process of feature extraction is the key to improving small target detection.

In order to improve the performance of small target detection under the roadside view and alleviate the large loss of fine-grained feature information during the convolution process, we introduced a novel convolutional neural network structure SPD-Conv [35] into the networks. SPD-Conv consists of a space-to-depth (SPD) layer and a non-strided convolution, which can completely replace the pooling and strided convolution layers in the traditional CNN module. Notably, the SPD layer downsamples the feature map X while preserving all the information in the channel dimension, thus avoiding the loss of information. As shown in Figure 6(a)-(c), applying SPD to an intermediate feature map X of size (S, S, C_1) yields a sequence of sub-feature maps:

$$\begin{aligned}
f_{0,0} &= X[0:S:scale, 0:S:scale],\\
f_{1,0} &= X[1:S:scale, 0:S:scale],\\
&\vdots\\
f_{scale-1,0} &= X[scale-1:S:scale, 0:S:scale];\\
f_{0,1} &= X[0:S:scale, 1:S:scale],\\
f_{1,1} &= X[1:S:scale, 1:S:scale],\\
&\vdots\\
f_{scale-1,1} &= X[scale-1:S:scale, 1:S:scale];\\
&\vdots\\
f_{scale-1,scale-1} &= X[scale-1:S:scale, scale-1:S:scale].
\end{aligned} \tag{3}$$

when scale $= 2$, four sub-feature maps $f_{0,0}, f_{0,1}, f_{1,0}$ and $f_{1,1}$ with size $(S/2, S/2, C_1)$ are obtained, by using downsampling operation on the feature map X. Then, they are concatenated to get a feature map X' with size $(S/2, S/2, 4C_1)$, and all the information in the channel dimension is preserved, thus no information is lost.

Finally, a non-strided convolution is added after the SPD module to reduce the information loss indiscriminately by increasing the use of learnable parameters in the convolution layers, as shown in Figure 6(d). In summary, the SPD-Conv module effectively reduces the loss of detailed features by adopting SPD instead of the traditional convolution for the downsampling operations. Therefore, the introduction of

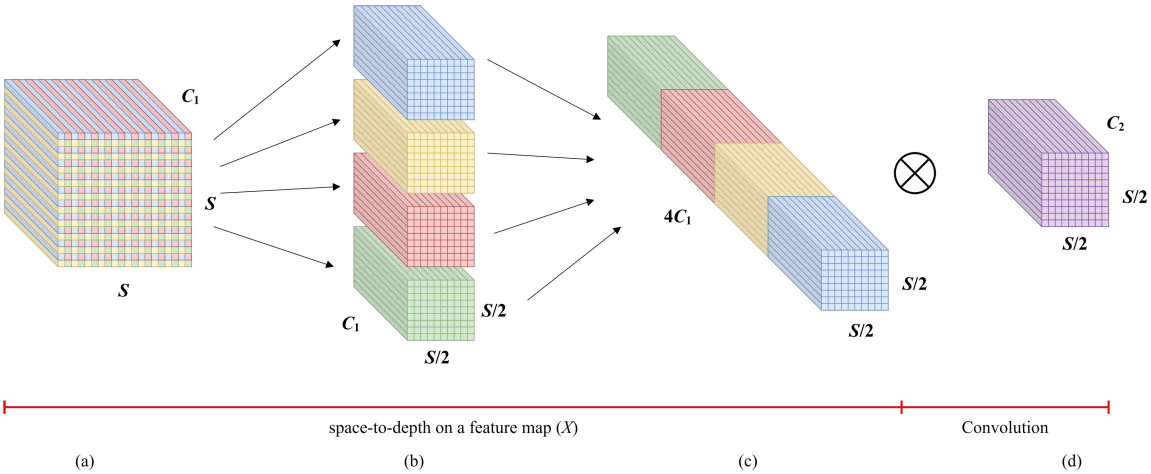


FIGURE 6. The overall framework of SPD-Conv. (a) Input feature map X. (b) Sub-feature maps $f_{0,0}$, $f_{0,1}$, $f_{1,0}$ and $f_{1,1}$ obtained by downsampling. (c) Feature map X′ is generated by concatenating sub-feature maps. (d) The output result is obtained after the convolutional dimensionality reduction operation.

SPD-Conv module can effectively improve the preservation rate of key features in low-resolution images and small target detection tasks, which is of great significance for improving the performance of roadside small target detection.

D. LIGHTWEIGHT FAST NEURAL NETWORK

Recently, with the deepening of deep neural network research, the development of computer vision technology has been greatly accelerated. Despite its impressive performance powering a range of applications, a major trend is to pursue fast neural networks with low latency and high throughput for better user experience, real-time responses, and security reasons, among others. This idea has been further validated recently on FasterNet [38].

Notably, the development of a fast neural network has extremely important for enhancing roadside target detection performance, improving driving safety, and developing intelligent transportation systems. As the computational resource of current roadside edge devices is limited, this demands a higher requirement for model lightweight. Additionally, the efficient collaboration of vehicle-road cooperative systems is dependent on effective data processing and transmission, thus the performance and efficiency of models are also of our concern. Currently, advanced anchor-free detectors, such as YOLOv6 and FCOS, have optimized the network architecture and performance. However, due to the adoption of large-capacity feature ex-traction backbones and feature extraction modules, the computational cost is evidently increased, making it difficult to directly apply to the real-time detection task under the circumstance of limited hardware resources in the roadside scenario.

In order to design a lightweight fast network model for roadside object detection, we introduce FasterNet Block into the YOLOv6 network to optimize the model performance. Notably, after being modified by FasterNet, our network has significantly reduced the number of parameters and computations without sacrificing the detection performance. Specifically, we mainly make improvements to the C3 module, replacing the Bottleneck with FasterNet Block. As shown in Figure 4, FasterNet Block consists of a PConv layer followed by two 1×1 Conv layers, with BN and ReLU applied after the intermediate Conv layer to achieve a balance between performance and speed.

As the core of FasterNet Block, PConv can reduce computation redundancy and memory access at the same time, as shown in Figure 5. Its core idea is to only apply filters to extract spatial features on a part of the input channels and keep the remaining channels unchanged. For continuous or regular memory access, the first or last continuous channels are taken as representatives of the entire feature map to calculate. Therefore, without sacrificing generality, assuming that the input and output feature maps have the same number of channels, the FLOPs of PConv is only $h \times w \times k^2 \times c_p^2$, which is equivalent to 1/16 of the conventional convolution. PConv also has a smaller amount of memory access, i.e., $h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p$, only 1/4 of the conventional convolution. In summary, we introduced FasterNet Block to replace the Bottleneck of C3, which can improve the efficiency of feature extraction while compressing the network volume.

E. TSCODE DECOUPLED DETECTION HEAD

Classification and localization are two highly related yet contradictory tasks in object detection. Classification is a coarse-grained task that requires a richer semantic context, whereas localization is a fine-grained task that requires more detailed boundary information [47]. Therefore, advanced detectors such as YOLOv6 and YOLOX have proposed a decoupled head to handle this conflict. Specifically, the output of the feature from the Neck is divided into two branches for classification and localization, respectively, and specific operations are performed in each task branch, as illustrated

in the formula.

$$\mathcal{T} = \mathcal{T}_{cls}(\mathcal{F}_c(N_l), \mathcal{C}) + \mathcal{T}_{loc}(\mathcal{F}_r(N_l), \mathcal{R}) \tag{4}$$

where $\mathcal{F}_c()$ and $\mathcal{F}_r()$ are the classification branch and localization branch, with the last layers $\mathcal{C}$ and $\mathcal{R}$ decoding the feature into classification scores and bounding box positions. $\mathcal{T}_{cls}$ and $\mathcal{T}_{loc}$ are the feature projection functions for classification and localization. In the common decoupling head designs, $\mathcal{T}_{cls}$ and $\mathcal{T}_{loc}$ have the same structure, but different parameters are provided for each task to provide different feature contexts, i.e. parameter decoupling.

However, such a simple design cannot fundamentally solve the problem. Because the semantics and spatial details information covered by different input features are not the same. Generally, low-level features have richer details but lack semantics, while high-level features are the opposite, which inevitably cannot maximize the advantages of the decoupled head. In addition, this design is largely determined by the input feature N_l, and the conflict between classification and localization leads to an imperfect balance between the two tasks.

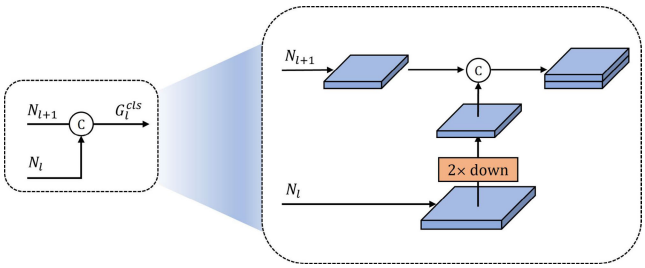


FIGURE 7. The classification branch focuses on semantic context encoding.

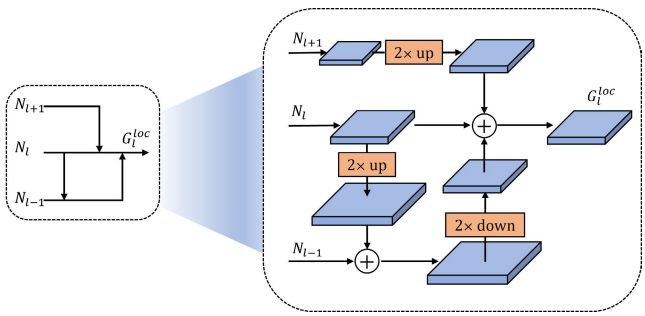


FIGURE 8. The localization branch focuses on detail-preserving encoding.

In order to maximize the performance of the decoupled head, we replaced the YOLOv6 decoupled head with the TSCODE head [47] to further improve the performance of roadside object detection. As shown in Figure 7, the classification task requires richer contextual semantic information and deep features can provide, so fusion from deep features can embed rich semantic information into the current feature map. On the other hand, the localization task needs richer spatial detail information, which shallow features can provide. Therefore, shallow features are guided back to the next layer feature map to achieve more reliable detection, as shown in Figure 8. It is worth mentioning that TSCODE does not use N_l as a common input, but provides task-specific input features G_l^{cls} and G_l^{loc} to the two branches, the principle of which is as follows:

$$G_l^{cls} = \text{Concat}(\text{Conv}(N_l), N_{l+1}) \tag{5}$$

$$G_l^{loc} = N_l + \mu(N_{l+1}) + \text{Conv}(\mu(N_l) + N_{l-1}) \tag{6}$$

$$\mathcal{T} = \mathcal{T}_{cls}(\mathcal{F}_c(G_l^{cls}), \mathcal{C}) + \lambda\mathcal{T}_{loc}(\mathcal{F}_r(G_l^{loc}), \mathcal{R}) \tag{7}$$

where Concat$(\cdot)$ represents channel concatenation operation; Conv$(\cdot)$ and $\mu(\cdot)$ represent a downsampling convolutional layer and upsampling. G_l^{cls} and G_l^{loc} with specific characteristics are fed to their respective task branches, thus maximizing the performance of the decoupled head.

IV. EXPERIMENT AND RESULTS

In order to verify the effectiveness of the PEFNet for roadside object detection, the model training and test experiment was conducted under the following working conditions: an Intel(R) Xeon(R) Platinum processor and 12GB running memory with the highest frequency of 2.50GHz; NVIDIA RTX 3080 GPU, 10GB graphics memory; Ubuntu operating system, CUDA version 11.0, Python version 3.8; deep learning framework PyTorch was used to establish roadside object detection model, and further improvement and optimization strategies were adopted. Finally, it was compared with the latest algorithms such as YOLOv6 and YOLOX on the Rope3D dataset and UA-DETRAC datasets.

A. EVALUATION METRICS

Before training the model, the training parameters of all the networks were set uniformly, with the sample batch size of 32, updating the weights once every 2 iterations, and the weight decay coefficient set to 0.0005. The training lasted for 150 iterations, with the initial learning rate (lr0) set to 0.01, the cycle learning rate (lrf) set to 0.1, and the learning rate momentum (momentum) set to 0.937.

Four metrics are employed in this model to measure the model's ability to recognize roadside targets, including precision (P), recall (R), category average precision (AP), and mean average precision (mAP). In addition, the model's detection speed is evaluated comprehensively by frames per second (FPS), giga floating-point operations per second (GFLOPs), and parameter count.

$$P = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

$$AP = \int_0^1 P(R)dR \tag{10}$$

$$mAP = \frac{1}{N}\int_0^1 P(R)dR \tag{11}$$

TABLE 1. Sample statistics of the Rope3D dataset based on size.

Dataset	Category	Images	Samples	Sample Amount		
				Small	Medium	Large
Training Set	Car	8,752	168,573	81,489	50,398	36,686
	Bus		10,205	4,472	2,116	3,617
	Van		14,225	5,686	4,420	4,119
	Truck		8,966	3,758	2,865	2,343
	Cyclist		20,674	11,881	6,641	2,152
	Tricyclist		8,671	4,940	2,127	1,604
	Pedestrian		28,508	20,053	7,633	822
	Motorcyclist		34,141	18,038	12,468	3,635
Test Set	Car	2,189	42,020	20,228	12,728	9,064
	Bus		2,540	1,103	514	923
	Van		3,346	1,358	994	994
	Truck		2,204	886	720	598
	Cyclist		5,484	3,153	1,779	552
	Tricyclist		2,152	1,217	528	407
	Pedestrian		7,208	5,019	1,953	236
	Motorcyclist		8,579	4,522	3,146	911

TABLE 2. Sample statistics of the UA-DETRAC dataset based on size.

Dataset	Category	Images	Samples	Sample Amount		
				Small	Medium	Large
Training Set	Car	3775	23,113	9,842	5,120	8,151
	Bus		1,543	50	115	1,378
	Van		2,616	847	537	1,232
	Others		163	27	26	110
Test Set	Car	329	2,037	801	472	764
	Bus		133	4	7	122
	Van		229	76	46	107
	Others		19	3	1	15

FIGURE 9. The samples of the datasets. (a) the samples of UA-DETRAC dataset (b) the samples of Rope3D dataset.

where true positives (TP) are the number of correctly predicted detection boxes of the roadside target category, false positives (FP) are the number of incorrectly recognized detection boxes of the roadside target category, false negatives (FN) are the number of undetected detection boxes of the roadside targets, i.e. the number of missed detections, and N is the number of categories.

B. DATASET ANALYSIS

In this paper, we selected the Rope3D [48] and UA-DETRAC [49] roadside view datasets for experiments, considering all factors. The Rope3D dataset was collected from various road scenes, including different lighting conditions (e.g. daytime, nighttime, dusk), weather conditions (e.g. rainy day, sunny day, cloudy day), and road environment. The UA-DETRAC is a public dataset from the field of autonomous driving, which is sourced from roadside monitoring videos and also includes different scenes and weather conditions. Generally, targets occupying less than 0.12% of the entire image are considered to be small targets, 0.12-0.5% are medium targets, and more than 0.5% are large targets. In the roadside object detection task, pedestrian targets are mostly small in size. The targets of cars are mostly small or medium. Targets of trucks and vans are mostly large. The datasets also include different illumination conditions and targets of different heights, and the shooting angle of the targets is also different.

Tables 1 and 2 show the sample statistics of our roadside dataset. It is evident that most targets are small or medium in size, and the target distribution is dense, which increases the difficulty of target detection. Figure 9 shows eight samples

TABLE 3. Performance comparison of the original Head and TSCODE Head on the Rope3D dataset.

Model	$mAP_{0.5}$ (%)	Params (10^6)	GFLOPs	FPS
YOLOv6	78.18	**17.19**	**44.08**	76.5
YOLOv6+TSCODE	**80.63**	17.36	50.11	**79.1**

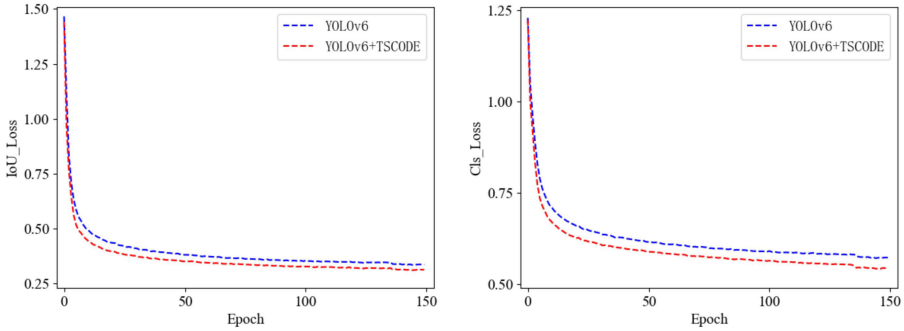


FIGURE 10. Loss-Curve comparison of the original and improved YOLOv6.

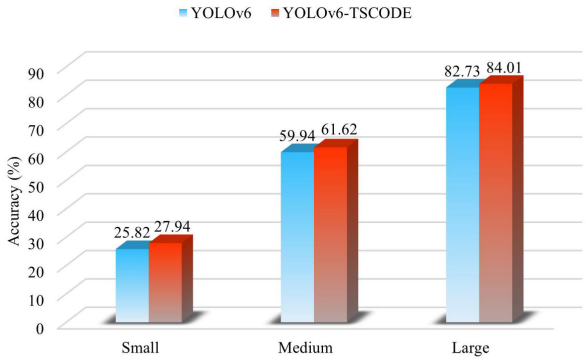


FIGURE 11. Accuracy comparison of original and improved YOLOv6 under different sizes.

from our dataset, where the targets reside in complex backgrounds. Further, after a series of convolution operations and downsampling layers, the targets occupy fewer pixels, thus making the detection more difficult.

C. PERFORMANCE COMPARISON ANALYSIS OF DETECTION HEAD

TSCODE head maximizes the performance of decoupled head by further subdividing classification and localization tasks. In order to verify the effectiveness of the improved detection head, loss function comparison experiments are conducted. During the experimental training process, we found that the loss curve tended to be stable when the epoch reached 150, so we terminated the training, and the loss results are shown in Figure 10.

From the comparison results of the above loss curves, it can be seen that as training epochs increase, the loss value gradually decreases and the loss curve tends to converge. When the epoch reached 150, the loss value was basically stable. Compared with the original YOLOv6, the improved YOLOv6 regression is faster and more accurate, demonstrating the effectiveness of the TSCODE Head. Moreover, from the training results in Table 3, the utilization of TSCODE shows better performance. Although a certain amount of computation is introduced, mAP0.50 are improved by 2.45% and the inference speed is almost unaffected.

To further illustrate the effect of TSCODE Head, we tested the detection performance of the model under different sizes, as shown in Figure 11. It is noteworthy that the model improved through TSCODE shows better performance in object detection at multiple scales, with an improvement in detection accuracy for small and medium-sized objects even exceeding 2%. Overall, the TSCODE Head proposed in this paper is effective for algorithm improvement.

D. EFFECTIVENESS OF THE PA-PAN STRUCTURE

To validate the effectiveness of the proposed PA-PAN structure, we built multiple object detection models with the same backbone network and detection head under the same conditions as the current advanced FPN structure. The results were verified on the Rope3D roadside dataset, as shown in Table 4.

The PANet achieved mAP0.50 of 78.18% for roadside object detection by integrating high-level semantics and low-level details through two feature transfer paths. Building upon PANet, GFPN added a cross-scale aggregation pathway that strengthened feature interactions and significantly improved model performance with a mAP0.50 of 81.29%. To decouple GFPN's complex structure, a more efficient ERepGFPN structure was proposed, which achieved nearly identical detection performance while greatly reducing model complexity and computational cost. Subsequently, RepBi-PAN was proposed, which further optimized model performance by fusing and preserving high-quality features, resulting in

TABLE 4. Performance comparison of the proposed FPN structure on the Rope3D dataset.

FPN Structures	AP (IoU=0.50:0.95) (%)			$mAP_{0.5}$ (%)	Params (10^6)	GFLOPs	FPS
	Small	Medium	Large				
PANet [46]	21.11	56.63	80.38	78.18	17.19	44.08	76.5
GFPN [29]	25.31	60.40	83.30	81.29	22.30	53.69	60.4
ERepGFPN [26]	25.41	60.02	82.58	81.20	15.98	42.91	**78.8**
RepBi-PAN [50]	26.36	59.83	82.64	81.39	17.86	45.74	75.3
PA-PAN (ours)	26.84	60.81	**83.72**	81.73	20.84	55.40	73.2
PA-PAN with FB (ours)	**28.03**	**61.42**	83.48	**82.27**	**10.32**	**34.92**	77.6

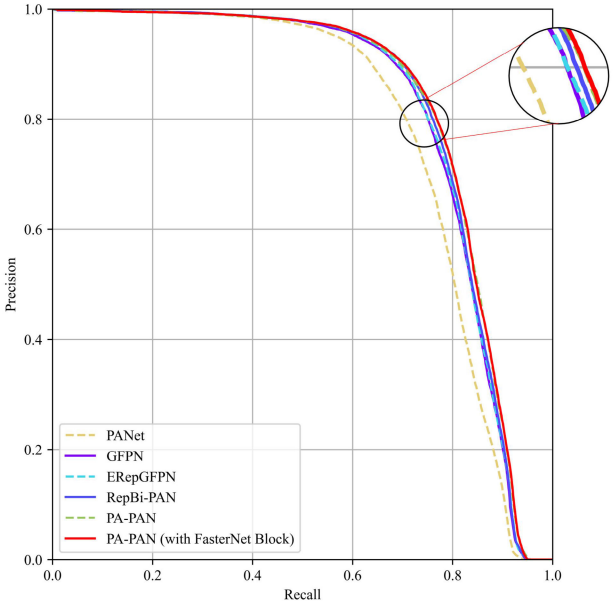


FIGURE 12. PR-Curve comparison of models composed of different FPN structures.

a mAP0.50 of 81.39%, higher than the previously proposed FPN structure, but with added computation.

Building upon the aforementioned research, and taking into account the practical environment at the roadside, we propose a novel PA-PAN that combines the advantages of ERep-GFPN and RepBi-PAN structures. In simple terms, we introduce a PEFA structure that enhances and aggregates shallow features with rich position information, making the network more focused on position information encoding. As shown in Table 4, the proposed PA-PAN structure achieves mAP0.50 of 81.73%, higher than the previously proposed FPN structure, proving the effectiveness of the PA-PAN architecture.

Although PA-PAN achieves better performance, its computational and parameter costs are not ideal, and even far exceed the previous FPN structure. Therefore, we introduce the FasterNet Block for lightweight optimization based on this, ultimately achieving the best performance. Furthermore, the PA-PAN with FasterNet Block (PA-PAN with FB) obtains the best AP value in detecting small objects, which further confirms that focusing on position information encoding can improve the detection performance of small objects.

To further validate the effectiveness of our proposed FPN structure, the PR-Curve comparison of models composed of different FPN structures is introduced. As shown in Figure 12, the PA-PAN with FasterNet Block achieves the best balance of accuracy and recall, which can be seen from the larger area under the curve. Based on the experimental results and analysis presented above, the proposed FPN structure can effectively improve the detection accuracy of small objects while maintaining model lightweight. These findings provide evidence that PA-PAN is an effective method for improving the performance of object detection models.

E. ABLATION STUDY AND ANALYSIS

In order to further validate the effectiveness of the proposed components, ablation experiments are conducted on the Rope3D dataset. In this paper, the ablation experiments are designed in two directions: (1) based on the original baseline algorithm, adding only one improvement strategy to verify the improvement effect of each strategy; (2) based on the final PEFNet algorithm, removing only one improvement strategy at a time to verify the effect of each strategy on the final algorithm.

In this experiment, only three improvement components were considered for the ablation experiments, as the SPD-Conv module has been integrated into the FasterNet Block and PA-PAN. The experimental results are shown in Table 5. Experiment 1 is the original YOLOv6 with mAP0.50 at 78.18%, FPS at 76.5, and Weight at 32.99MB. In Experiment 2, based on the original YOLOv6, a lightweight network architecture was constructed using the FasterNet Block. The Params and GFLOPs increased by 63.1% and 50.3%, respectively, indicating that the FasterNet Block can effectively reduce the model volume and complexity. In Experiment 3, PA-PAN was adopted to replace the original PANet for feature fusion. Although the model introduced additional calculations, it effectively improved the accuracy of small object detection. The mAP0.50 value increased by 3.55%. In Experiment 4, the TSCODE module was aggregated into the Head, demonstrating that decoupling the classification and localization tasks can further improve the performance of the detection head.

Up to this point, all the improved modules have achieved better improvement based on the original YOLOv6 model,

TABLE 5. Comparison of ablation experiment results.

NO	FasterNet Block	PA-PAN	TSCODE Head	$mAP_{0.50}$ (%)	Params (10^6)	GFLOPs	FPS	Weight (MB)
1				78.18	17.19	44.08	76.5	32.99
2	✓			80.15	**6.35**	**21.92**	**81.8**	**12.39**
3		✓		81.73	20.84	55.40	72.3	39.99
4			✓	80.63	17.36	50.11	79.1	33.34
5	✓	✓		82.27	10.32	34.92	77.6	20.01
6	✓		✓	80.42	6.52	27.96	80.8	12.74
7		✓	✓	82.11	21.01	61.44	70.9	40.34
8	✓	✓	✓	**82.39**	9.71	39.00	75.4	18.80

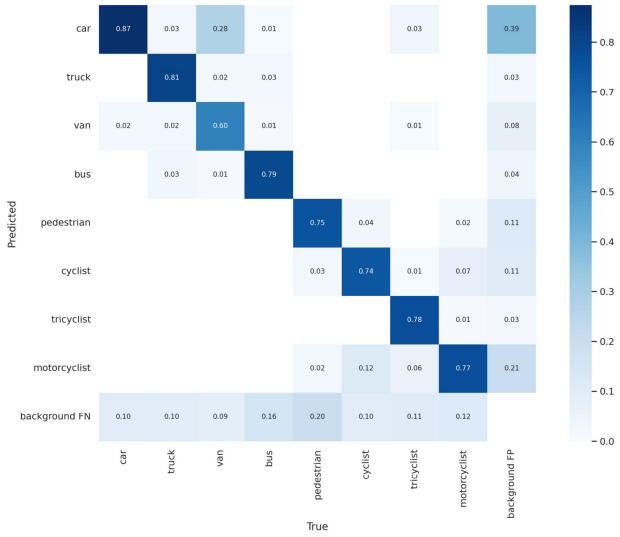


FIGURE 13. Confusion Matrix representation of PEFNet on Rope3D dataset.

but there is still great potential for improvement. Therefore, we further integrate multiple improved modules to achieve better performance. In Experiment 5, FasterNet Block and PA-PAN were integrated into the original YOLOv6 model, compared with Experiment 3 which only introduced PA-PAN, the mAP0.5 was improved by 0.54%. Meanwhile, the Params and GFLOPs were decreased significantly, and the model size compression effect was obvious. In Experiment 6, based on the lightweight model improved by FasterNet Block, TSCODE was added to enhance the performance of the detection head. mAP0.5 increased by 0.27%. Similarly, in experiment 7, TSCODE was integrated into the model improved by PA-PAN. mAP0.5 increased by 0.38%. This further demonstrates that the improved detection head can provide better performance.

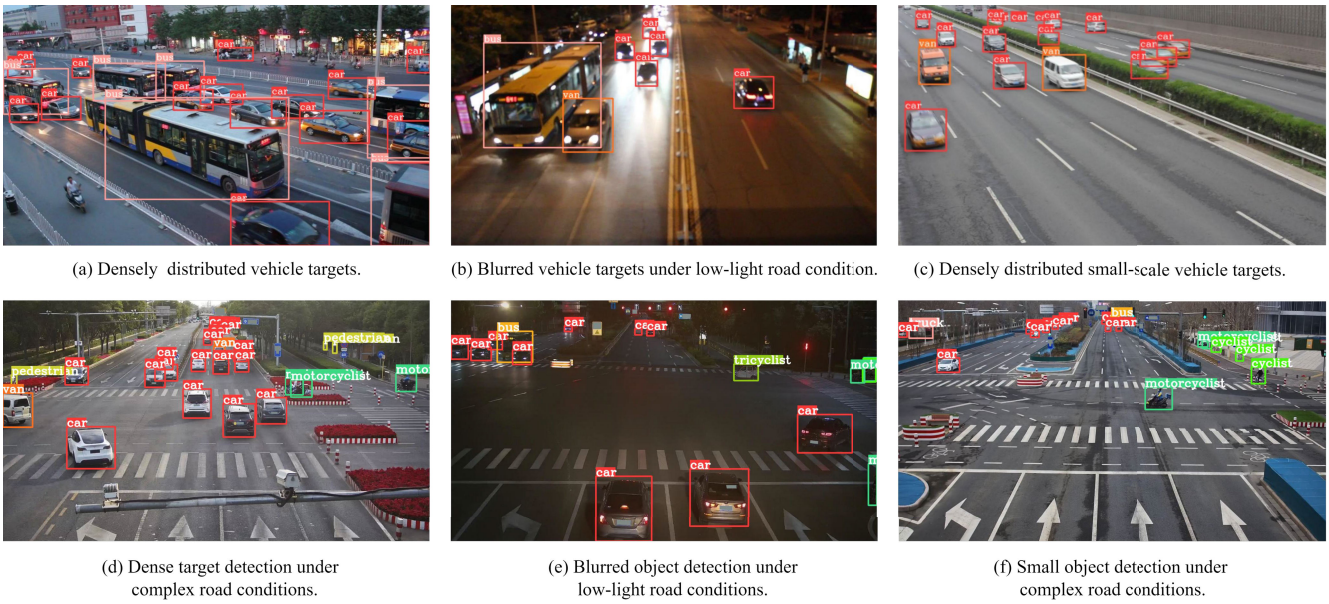
Finally, all the improved modules are combined to achieve the best performance of the detection model and named the model PEFNet. Compared with the original YOLOv6 model, the mAP0.50 value increased by 4.21% and the model weight decreased by 43.1%, while the detection speed can be maintained at 75.4fps. This shows that the detection performance of PEFNet is better than YOLOv6, further confirming the effectiveness of the three improvement methods mentioned above. The confusion matrix of PEFNet is shown in Figure 13. It can be observed that PEFNet exhibits significant improvements in both accuracy and recall, along with a clear reduction in error rates, indicating that our improved model is effective.

F. EXPERIMENT RESULTS AND ANALYSIS ON ROPE3D AND UA-DETRAC DATASETS

In order to further verify the effectiveness of the proposed algorithm, mAP, and FPS metrics are chosen to evaluate the accuracy and real-time performance, respectively. Our algorithm is compared with the advanced object detection models on the Rope3D dataset, and the comparison results are shown in Table 6. The results of different sizes are shown in Table 7.

As shown in Table 6, the proposed PEFNet achieved 82.39% and 58.34% in mAP0.50 and mAP0.95, respectively, which is significantly better than FCOS, YOLOX, YOLOv3, and YOLOv6. In comparison, YOLOv6v3.0 and YOLOv7s can provide detection results close to PEFNet. However, compared with the well-performing YOLOv6v3.0, our proposed PEFNet reduces Params and GFLOPs by 45.6% and 14.7%, respectively. In particular, the weight of PEFNet decreased by 51.3%, the model complexity decreased significantly. In addition, as shown in the detection results of different-sized targets in Table 7, for small-scale targets primarily consisting of pedestrians and cyclists, our proposed method outperforms all over algorithms in terms of detection accuracy, with a lower leak detection rate. Overall, compared to the original YOLOv6, our algorithm has achieved significant improvements in model performance and lightweight characteristics, particularly with its significant advantages in small object detection.

In order to ensure the generalizability of the algorithm, we conducted experiments on the UA-DETRAC dataset. The comparison results are shown in Table 8. Compared with other network models, PEFNet shows more powerful performance in both detection accuracy and model lightweight, especially in small target detection, PEFNet significantly outperforms other networks, which further proves the effectiveness of our proposed PA-PAN feature fusion structure. Furthermore, it can be seen from Table 8 that the leak detection rate of the algorithm is lower than that of the advanced



(a) Densely distributed vehicle targets. (b) Blurred vehicle targets under low-light road condition. (c) Densely distributed small-scale vehicle targets.

(d) Dense target detection under complex road conditions. (e) Blurred object detection under low-light road conditions. (f) Small object detection under complex road conditions.

FIGURE 14. The detection results of the PEFNet on UA-DETRAC and Rope3D datasets.

TABLE 6. Comparison among PEFNet and current advanced detectors in terms of detection accuracy and efficiency on the Rope3D dataset.

Model	Backbone	Metric					FPS
		$mAP_{0.50}$ (%)	$mAP_{0.95}$ (%)	Params (10^6)	GFLOPs	Weight (MB)	
YOLOv3 [20]	DarkNet53	75.69	41.61	61.56	77.62	118.75	44.3
FCOS [12]	ResNet50	77.34	50.39	31.85	78.76	54.56	25.3
YOLOX [21]	CSPDarkNet	75.40	49.83	**8.89**	**13.33**	**15.56**	42.1
YOLOv6 [25]	EfficientRep	78.18	53.23	17.19	44.08	32.99	76.5
YOLOv6v3.0 [50]	EfficientRep	81.39	56.71	17.86	45.74	38.70	75.3
YOLOv7s [22]	ELAN-Net	81.69	55.54	9.15	26.13	18.13	**76.9**
PEFNet (ours)	**FasterNet**	**82.39**	**58.34**	9.71	39.00	18.80	75.4

TABLE 7. Experimental comparisons of accuracy measured by size on the Rope3D dataset.

Model	Backbone	AP (IoU=0.50:0.95) (%)			Leak Detection Rate (%)
		Small	Medium	Large	
YOLOv3 [20]	DarkNet53	19.12	46.22	59.71	21.6
FCOS [12]	ResNet50	21.51	54.41	72.73	16.0
YOLOX [21]	CSPDarkNet	21.61	51.93	75.92	15.8
YOLOv6 [25]	EfficientRep	21.11	56.63	80.38	13.4
YOLOv6v3.0 [50]	EfficientRep	26.36	59.83	82.64	11.9
YOLOv7s [22]	ELAN-Net	25.42	58.64	81.62	14.3
PEFNet (ours)	**FasterNet**	**28.12**	**60.00**	**83.22**	**11.4**

algorithms. Therefore, it can be concluded that the proposed PEFNet can provide better roadside detection results.

Finally, under different backgrounds, our model is tested on partial images of the UA-DETRAC and Rope3D datasets to obtain visualization results. As shown in Figure 14, the visualization results show that PEFNet can accurately detect targets under different illumination, distribution, and size conditions, and it exhibits superior detection performance for multi-scale roadside targets. Overall, PEFNet demonstrated good performance in roadside object detection tasks, as it almost detected all targets and classified them correctly in various road scenarios, showing its robust generalization ability.

G. VISUALIZATION RESULTS AND ANALYSIS ON ROPE3D AND UA-DETRAC DATASETS

To validate the performance of the proposed network in cross-scene detection tasks, we further explore a dataset of roadside environments with various complex scenarios to evaluate the network's robustness to scenario changes. Figure 15 shows the qualitative test results of FCOS, YOLOX, YOLOv6, and PEFNet for roadside targets under different scenarios and

TABLE 8. Comparison among PEFNet and current advanced detectors in terms of accuracy and speed on the UA-DETRAC dataset.

Model	Backbone	Metric (%)					FPS
		Small	Medium	Large	$mAP_{0.50}$ (%)	Leak Detection Rate	
YOLOv3 [20]	DarkNet53	16.22	58.21	55.14	88.12	10.3	43.8
FCOS [12]	ResNet50	19.63	66.23	71.26	86.53	9.0	27.1
YOLOX [21]	CSPDarkNet	17.10	66.92	78.95	90.15	3.0	68.4
YOLOv6 [25]	EfficientRep	21.94	70.35	82.91	92.84	2.8	73.1
YOLOv6v3.0 [50]	EfficientRep	23.36	72.54	83.96	93.63	2.0	**79.8**
YOLOv7s [22]	ELAN-Net	23.96	73.81	84.93	94.23	2.0	71.4
PEFNet (ours)	**FasterNet**	**24.81**	**76.93**	**85.12**	**95.61**	**1.0**	73.9

FIGURE 15. Comparison of roadside detection in different scenarios captured by roadside visual sensors (Rope3D dataset). Cases of missed detection and false detection are highlighted in red.

weather conditions based on the Rope3D dataset. Column a shows roadside images taken in cloudy environments, column b shows roadside images taken in night environments, column c shows roadside images taken in rainy environments, and column d shows roadside images taken in sunny environments. These images contain a large number of small and blurred targets, making them more difficult to detect. Under the influence of illumination and background noise, targets with less obvious features such as pedestrians are more likely to be missed or falsely detected.

Based on the analysis of the graph, PEFNet shows better detection performance compared to other network algorithms, with fewer missed and false detections. Especially in well-lit environments, the improvements in roadside object detection are more evident. As shown in column a and column c of Figure 15, YOLOv6, YOLOX, and FCOS performed poorly in detecting small targets, whereas PEFNet was able to detect these small targets more effectively and classify them correctly. It's worth noting that our model also impressed in scenarios with high background noise interference. As shown in column d of Figure 15, YOLOv6, YOLOX, and FCOS all exhibited false detections and missed detections for targets that were difficult to identify in shadows. However, PEFNet not only accurately detected the targets but also correctly classified them. This indicates that PEFNet can effectively reduce the interference of background noise, thus identifying the target more accurately. Additionally, column b of Figure 15 displays detection results in low-light scenes that contain more blurry and small-scale targets. Compared with other models, PEFNet can better recognize low-light targets and locate small-scale targets far away. Overall, compared to the original YOLOv6, PEFNet has significantly improved

FIGURE 16. Comparison of roadside detection in different scenarios captured by roadside visual sensors (UA-DETRAC dataset). Cases of missed detection and false detection are highlighted in blue.

the detection capability, with both false negative and false positive rates greatly reduced.

To further verify the generalization performance of our proposed algorithm, we extracted multiple sets of complex scene roadside images from the UA-DETRAC dataset for validation. As shown in column a and column b of Figure 16 in well-lit environments, YOLOv6, YOLOX, and FCOS all exhibited missed and false detections, whereas PEFNet accurately detected all targets and categorized them correctly. Additionally, as shown in the detection results of column c and column d of Figure 16 in low-light environments, our proposed model achieved higher accuracy compared to YOLOv6, YOLOX, and FCOS and was more sensitive to small targets. In summary, PEFNet showed lower false positives and false negatives in dense object environments and complex scenes (such as dimly lit nights).

In conclusion, due to the effective design and novel structure of the network, PEFNet has sufficient accuracy and robustness improvement to perform cross-scene detection efficiently and accurately. Furthermore, PEFNet can also achieve good results on different datasets, which not only verifies its generalization ability but also brings potential promotion to future intelligent transportation development.

V. CONCLUSION

This paper proposes a novel roadside image target detection algorithm PEFNet based on YOLOv6. It aims to address the challenge of balancing detection speed and accuracy in roadside object detection due to small target size, complex background, and limited feature extraction capabilities. To achieve this, the position-aware feature pyramid network is proposed to improve small object detection performance and multi-scale feature generalization ability. The FasterNet Block is introduced into the Backbone and Neck networks to provide efficient feature extraction while achieving network lightweight transformation. The SDP-Conv is inserted in the network to enhance effective feature extraction. Furthermore, the TSCODE is aggregated into the detection head to achieve accurate target recognition and suppress background noise interference. The experimental results on the Rope3D and UA-DETRAC datasets show that PEFNet achieves 4.21% and 2.77% mAP improvements, respectively, compared to the original YOLOv6. By focusing on position information coding, PEFNet improves the detection of small targets, with the AP of small objects increasing by 7.01% on the Rope3D dataset and 2.87% on the UA-DETRAC dataset. In addition, by introducing FasterNet Block lightweight structure, the model volume is reduced by 43.1%. Meanwhile, our method maintains a detection speed of 75fps and obtains higher accuracy than current advanced detection algorithms. Overall, our method achieves satisfactory performance in roadside object detection while maintaining real-time capability.

Compared with anchor-based detectors, anchor-free detectors still have great potential for optimization in terms of accuracy and performance. However, in the field of roadside object detection, accuracy improvement is still restricted by

the complex environment, large-scale variance, and rotation change. In the future, we need to focus not only on model structure and paradigm but also on improving the robustness and stability of models to achieve more reliable detection. We will continue to adjust the hyperparameters, optimize the model, and further improve the speed and accuracy of roadside object detection.

REFERENCES

[1] M. Tsukada, T. Oi, M. Kitazawa, and H. Esaki, "Networked roadside perception units for autonomous driving," *Sensors*, vol. 20, no. 18, p. 5320, Sep. 2020.

[2] A. Chtourou, P. Merdrignac, and O. Shagdar, "Collective perception service for connected vehicles and roadside infrastructure," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–5.

[3] Z. Bai, G. Wu, X. Qi, Y. Liu, K. Oguchi, and M. J. Barth, "Infrastructure-based object detection and tracking for cooperative driving automation: A survey," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 1366–1373.

[4] P. Sun, C. Sun, R. Wang, and X. Zhao, "Object detection based on roadside LiDAR for cooperative driving automation: A review," *Sensors*, vol. 22, no. 23, p. 9316, Nov. 2022.

[5] L. Huang and W. Huang, "RD-YOLO: An effective and efficient object detector for roadside perception system," *Sensors*, vol. 22, no. 21, p. 8097, Oct. 2022.

[6] Z. Zhang, J. Zheng, H. Xu, X. Wang, X. Fan, and R. Chen, "Automatic background construction and object detection based on roadside LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4086–4097, Oct. 2020.

[7] Z. Xu, S. Zhao, and R. Zhang, "An efficient multi-sensor fusion and tracking protocol in a vehicle-road collaborative system," *IET Commun.*, vol. 15, no. 18, pp. 2330–2341, Nov. 2021.

[8] H. Jiao, "Intelligent research based on deep learning recognition method in vehicle-road cooperative information interaction system," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Jun. 2022.

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

[10] 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.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[11] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.

[12] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.

[13] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[14] C. Wang, H. M. Liao, Y. Wu, P. Chen, J. Hsieh, and I. 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. 1571–1580.

[15] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[16] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An enhanced MobileNet architecture," in *Proc. IEEE 10th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2019, pp. 0280–0285.

[17] X. Chen, L. Ma, J. Chen, Z. Jie, W. Liu, and J. Luo, "Real-time referring expression comprehension by single-stage grounding network," 2018, *arXiv:1812.03426*.

[18] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4682–4692.

[19] Y. Liao, S. Liu, G. Li, F. Wang, Y. Chen, C. Qian, and B. Li, "A real-time cross-modality correlation filtering method for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10877–10886.

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

[21] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[22] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 7464–7475.

[23] 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. 13th Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 740–755.

[24] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.

[25] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[26] X. Xu, Y. Jiang, W. Chen, Y. Huang, Y. Zhang, and X. Sun, "DAMO-YOLO: A report on real-time object detection design," 2022, *arXiv:2211.15444*.

[27] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13728–13737.

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

[29] Y. Jiang, Z. Tan, J. Wang, X. Sun, M. Lin, and H. Li, "GiraffeDet: A heavy-neck paradigm for object detection," 2022, *arXiv:2202.04256*.

[30] A. Li, S. Sun, Z. Zhang, M. Feng, C. Wu, and W. Li, "A multi-scale traffic object detection algorithm for road scenes based on improved YOLOv5," *Electronics*, vol. 12, no. 4, p. 878, Feb. 2023.

[31] L. Jiang, B. Yuan, W. Ma, and Y. Wang, "JujubeNet: A high-precision lightweight jujube surface defect classification network with an attention mechanism," *Frontiers Plant Sci.*, vol. 13, Jan. 2023, Art. no. 1108437.

[32] J. Du, K. Yang, Y. Hu, and L. Jiang, "NIDS-CNNLSTM: Network intrusion detection classification model based on deep learning," *IEEE Access*, vol. 11, pp. 24808–24821, 2023.

[33] X. Shao, C. Wei, Y. Shen, and Z. Wang, "Feature enhancement based on CycleGAN for nighttime vehicle detection," *IEEE Access*, vol. 9, pp. 849–859, 2021.

[34] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020.

[35] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*, Grenoble, France. Cham, Switzerland: Springer, 2023, pp. 443–459.

[36] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[37] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, and E. Chi, "DCN v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems," in *Proc. Web Conf.*, Apr. 2021, pp. 1785–1797.

[38] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher FLOPS for faster neural networks," 2023, *arXiv:2303.03667*.

[39] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

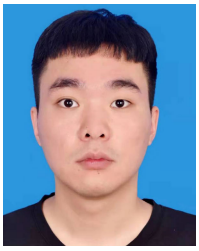
[40] Q. Wu, J. Wang, Z. Chai, and G. Guo, "Multi-scale feature aggregation and boundary awareness network for salient object detection," *Image Vis. Comput.*, vol. 122, Jun. 2022, Art. no. 104442.

[41] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1968–1979, 2022.

[42] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018.

[43] 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. 91–99.

[44] 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.
[45] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[46] 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.
[47] J. Zhuang, Z. Qin, H. Yu, and X. Chen, "Task-specific context decoupling for object detection," 2023, *arXiv:2303.01047*.
[48] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, and E. Ding, "Rope3D: The roadside perception dataset for autonomous driving and monocular 3D object detection task," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21309–21318.
[49] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102907.
[50] C. Li, L. Li, Y. Geng, H. Jiang, M. Cheng, B. Zhang, Z. Ke, X. Xu, and X. Chu, "YOLOv6 v3.0: A full-scale reloading," 2023, *arXiv:2301.05586*.



LEI HUANG received the B.E. degree from Jiaxing Nanhu University, Jiaxing, Zhejiang, China, in 2021. He is currently pursuing the M.E. degree with Xijing University, Xi'an, China. His current research interests include deep learning and computer vision.

WENZHUN HUANG received the B.S. and M.S. degrees in communication and information systems from Air Force Engineering University, China, in 1994 and 1997, respectively, and the Ph.D. degree in information and communication engineering from Northwest Polytechnic University, China, in 2010. He is currently a Professor at Xijing University. His research interests include information processing, big data, wireless communication systems, and the IoT technology.

HAI GONG received the B.E. degree from the Polytechnic Institute of Jiangxi Science and Technology Normal University, Jiangxi, China, in 2021. He is currently pursuing the M.E. degree with Xijing University, Xi'an, China. His current research interests include deep learning and intelligent information processing.

CHANGQING YU is a Professor at Xijing University, Xi'an, China. His research interests include machine learning, neural networks, data mining, data analysis, and bioinformatics.

ZHUHONG YOU is a Professor at Northwestern Polytechnical University, Xi'an, China. His research interests include neural networks, intelligent information processing, sparse representation, and its applications in bioinformatics.