IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 October 2023, accepted 17 November 2023, date of publication 20 November 2023,
date of current version 28 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335230

RESEARCH ARTICLE

Lightweight Real-Time Detection and Recognition Model of Intraocular Foreign Bodies Fused With a Feature Pyramid Mechanism

YIRAN LIU[1,2], YITING ZHENG[3], XIAOYU ZHU[3], JUNZHE CHEN[3], SUYAN LI[2], AND ZHAOLIN LU[2]

[1]The Second Clinical Medical College, Nanjing Medical University, Nanjing, Jiangsu 211166, China
[2]Xuzhou No. 1 People's Hospital, Xuzhou, Jiangsu 221002, China
[3]School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China

Corresponding author: Zhaolin Lu (zhaolinlu@cumt.edu.cn)

This work was supported in part by the National Natural Science Foundations of China under Grant 62371451, in part by the Xuzhou City Social Development Key Special Project under Grant KC21153, in part by the Hospital Management Innovation Research Project of Jiangsu Hospital Association under Grant JSYGY-3-2023-353, and in part by the Xuzhou Science and Technology Project under Grant KC21262.

ABSTRACT Accurate detection of target location and type is crucial for treating ocular trauma caused by foreign bodies intrusion. However, the traditional method of manually marking CT image targets has slow recognition speed and poor detection accuracy, which cannot meet the real-time and accuracy requirements for detecting foreign bodies in clinical diagnosis. To address this issue, we propose a lightweight detection and recognition model based on feature extraction and fusion. Firstly, the normalization-based attention module and the sigmoid linear unit activation function are introduced into the inverted residual block of the backbone network to enhance the model's attention to salient features and improve the detection accuracy. Then, the path aggregation feature pyramid network is utilized to fuse multiscale features, enabling the information interaction between different levels of the network and enhancing the accuracy foreign bodies classification. In particular, the incorporation of the space-to-depth convolution and convolutional mixing modules into the feature pyramid network significantly reduce the computational overhead while effectively capturing the key semantic features in both space and channel directions, thereby improving the lightweight level of the model. Finally, the location and type information of the foreign intraocular bodies are obtained by this model. The experimental results demonstrate the superior performance of the proposed model in terms of mAP@0.5, accuracy, sensitivity and specificity, achieving 97.2, 93.5, 98.0 and 88.0, respectively. Furthermore, the smaller number of parameters and faster detection time allow the proposed model run in real-time on poorly configured hardware, making it more suitable for clinical applications.

INDEX TERMS Intraocular foreign body, feature pyramid, lightweight, real-time detection.

I. INTRODUCTION

Foreign bodies invasion of the eye refers to the entry of various foreign bodies into the eye through the eyewall, which is a serious ocular trauma with a high risk of blindness [1]. Due to the inherent fragility of ocular tissues, including the lens and retina, natural recovery is impossible [2]. Therefore, prompt and precise diagnosis of the injury location and type is essential [3]. Different intraocular foreign bodies can cause varying levels of damage to the eye, leading to a need for different treatment methods. Hence, the location and type information of intraocular foreign bodies are crucial in determining the surgical approach and achieving successful treatment outcomes.

Medical imaging examination is currently the core means of clinical disease screening, diagnosis, treatment and

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

evaluation, among which CT scanning is the simplest and non-invasive and [4] less harmful to the human body compared to traditional radiography. Furthermore, CT scanning is highly sensitive to soft tissues and organs, allowing for clear visualization of lesion sites and obtaining high diagnostic value images. As a result, it is widely utilized for examining intraocular foreign bodies. Currently, the methods for detecting foreign bodies are still at the stage of manually marking CT images [5]. The accuracy and efficiency in determining the orientation and type of foreign body does not meet the needs of clinical examination [3], [6]. Therefore, there is an urgent need for an automatic image detection and recognition method.

With the development of artificial intelligence(AI) technology [7], [8], image processing methods based on deep learning(DL) are widely used in the medical field [9], [10], [11]. For example, Khan et al. [12] proposed a pyramid-based multiscale encoder-decoder network for medical image segmentation that demonstrated model segmentation accuracy and diagnostic speed on four publicly available medical datasets. Zhao et al. [13] used a DL model combined with migration learning and feature fusion techniques to assist physicians in the clinical diagnosis of thyroid nodules, which improves diagnostic accuracy and efficiency. Especially in the field of ophthalmology, many researchers have used medical images to build DL models to analyze various ophthalmic diseases [14]. Pan et al. [15] employed a DL algorithm to automatically detect four lesions in classified fluorescein images, achieving acute DR grading. Lin et al. [16] introduced bounded heat map regression and signed distance map reconstruction branches on top of the segmentation framework, and proposed the first joint learning framework, BSDA-Net, for centro-concave avascular zone segmentation and multi-disease classification. To the best of our knowledge, there are no AI methods for detecting and recognizing intraocular foreign bodies.

Furthermore, there is a growing tendency in DL applications to implement models on end-side platforms, such as mobile and embedded devices, that are capable of running in real-time and in real-world environments. These platforms are distinguished by their low memory and processor capabilities, which impedes the deployment of networks requiring high memory and computational resources, necessitating the development of lightweight network architectures. SqueezeNet [17] is an early and classic lightweight network that uses the Fire module for parameter compression. The MobileNet [18], [19], [20], [21]series network is the predecessor of the MobileViT [22] series network, which has a smaller volume, fewer calculations, higher accuracy and great advantages in lightweight neural networks. A fully connected layer-based attention mechanism was proposed in the GhostNetv2 [23] architecture, which possesses the ability to quickly execute on common hardware whilst capturing dependencies between long-distance pixels. ShuffleNetv2 [24] addressed the issue of inadequate feature channels in the face of limited computing resources. The You Only Look Once(YOLO) [25], [26], [27] family is a typical class of one-stage target detection algorithms that use anchor boxes to combine classification with the regression issue of target location, leading to high efficiency, flexibility and good generalisation performance.

At present, there exist three primary challenges in the intelligent detection and recognition of intraocular foreign bodies in CT images. Firstly, the area of the foreign body part is extremely small compared with the overall brain CT image, resulting in a high missed detection rate. Secondly, some structures in the brain regions external to the eyeball are highly similar to intraocular foreign bodies, leading to false detection of foreign bodies. Lastly,the detection times of traditional algorithms is slow and cannot meet the actual needs of rapid clinical diagnosis. To address these issues and meet the need for lightweight, we propose a real-time method for detecting and recognising intraocular foreign bodies, which uses MobileViTv3 network as the feature extraction model and path aggregation feature pyramid network (PAFPN) as the feature fusion model. The proposed method can obtain the location and type information of the foreign bodies in real time. Compared with other lightweight models, the proposed model has lower computational overhead, less detection times and higher detection accuracy.

Therefore, to address these issues and meet the need for lightweight, we propose a real-time method for the detection and recognition of intraocular foreign bodies, which uses MobileViTv3 network as the feature extraction model and path aggregation feature pyramid network (PAFPN) as the feature fusion model. The proposed method can obtain the location and type information of the foreign bodies in real time.

II. METHODS

From the perspective of functional structure, the proposed model mainly includes two parts: feature extraction and feature fusion. Among them, MobileViTv3 is used as the backbone network for extracting features, and PAFPN is used for fusing features. Taking the eye CT image as input, the location and type information of the intraocular foreign bodies are obtained after feature extraction and fusion processing, as shown in Figure 1. In the following, the detailed architecture of the feature extraction and fusion model will be introduced in detail.

A. FEATURE EXTRACTION MODEL

The backbone network is mainly composed of multiple inverted residual block(IRB) and multiscale representation block(MRB), which are used to extract local and global visual features of CT images. In order to enhance the model's attention to salient features of foreign intraocular bodies and improve the calculation accuracy of the deep network, the normalization-based attention module (NAM) and SiLU activation function are introduced into the IRB. Then, three feature layers of different scales are output by three MRBs as the input of the feature fusion network.

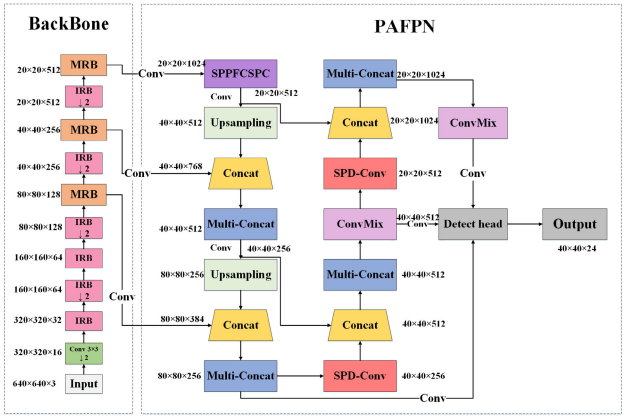

FIGURE 1. Block diagram of detection and recognition model fused with feature pyramid.

1) IRB MODULE

Different from the traditional residual block, the IRB structure has a large number of intermediate channels but a small number of channels at both ends. The purpose is to adopt the strategy of initially increasing the number of channels and then reducing the number of channels to balance the contradiction between the number of convolution calculations and the number of feature channels. The IRB module extends the input features to higher dimensions to increase the nonlinear changes in each channel to obtain higher detection accuracy with less calculation. Figure 2 shows the three parts of the IRB module.

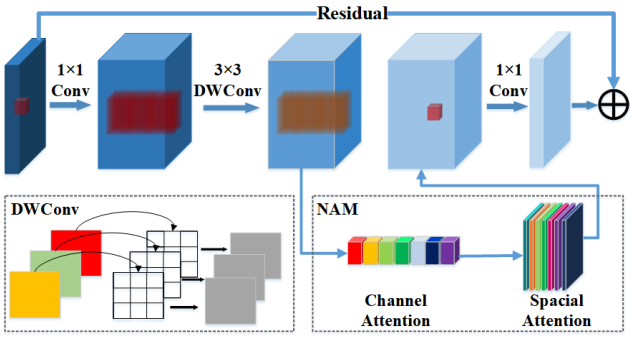

FIGURE 2. IRB module.

(1) In the first part, a Conv operation of convolution of size 1×1 is performed on the input feature layer to achieve dimensionality elevation. Then, the batch normalisation (BN) layer is used to normalise the features, which makes the loss function smoother and conducive to gradient descent. In this study, the sigmoid linear unit(SiLU) activation function is used as shown in formula 1, which is better than the original activation function and has stronger robustness under low-precision calculation.

$$SiLU(x) = x\sigma(x), \sigma = \frac{1}{1+e^{-x}} \tag{1}$$

(2) In the second part, a 3×3 depthwise convolution (DWConv) module is used to extract features, which is also connected with the BN layer and SiLU activation function. As shown in the lower left of Figure 2, the channel of each convolution kernel of DWConv is 1; it is only responsible for one channel of the input feature matrix. Therefore, the number of convolution kernels must be equal to the number of channels of the input feature matrix. The number of channels of the convolution outputfeature matrix is also equal to the number ofchannels of the input feature matrix.

(3) In the third part, dimension reduction is achieved using the Conv convolution of size 1×1, followed by the BN processing layer. The IRB structure removes the final activation function in this part. When the input dimension is the same as the output dimension, the IRB structure connects the output with the input residual, which avoids the gradient failure phenomenon during the depth convolution process. The single-depth convolution does not increase the number of parameters in the high-dimensional feature layer.

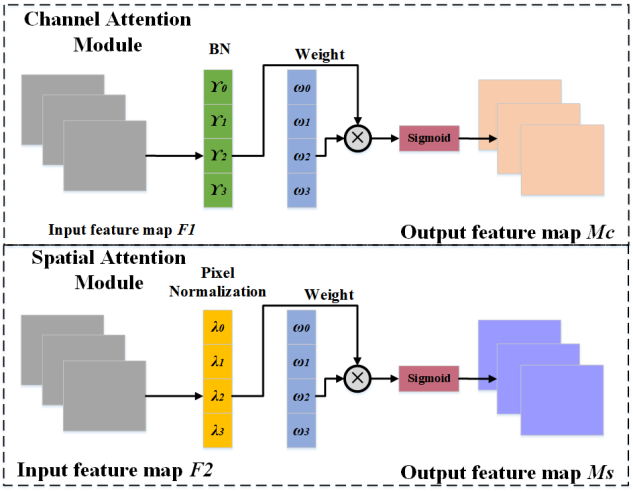

FIGURE 3. NAM module.

In this study, the NAM module is added before feature dimension reduction. Furthermore, the contribution factor of weight is used to improve the performance of the attention mechanism. The NAM adopts the integration method of the channel-spatial hybrid attention module and redesigns the channel attention and spatial attention submodules. For the channel attention submodule, the scaling factor in BN is used, and the channel variance is calculated by the scaling factor as the basis for measuring the importance of weights:

$$B_{\text{out}} = BN(B_{\text{in}}) = \gamma \frac{B_{\text{in}} - \mu_\beta}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{2}$$

where μ_β and σ_β denote the mean and standard deviation of the minibatch β, respectively, and γ and β are trainable affine transformation parameters. The channel attention and spatial attention modules in the NAM are shown in Figure 3, where M_c represents the output features of the channel attention

module, as shown in Eq(3), and γ_i represents the scaling factor of each channel. The NAM also applies the BN scaling factor to the spatial dimension to measure the importance of pixels, which is called pixel normalisation. The spatial attention module in the NAM is shown in the Eq (4),where M_s represents the output of the spatial attention module and λ_i represents the scaling factor.

$$M_c = \text{Sigmoid}\left(W_\gamma\left(BN\left(F_1\right)\right)\right), \quad W_\gamma = \gamma_i / \sum_{j=0} \gamma_j \tag{3}$$

$$M_s = \text{Sigmoid}\left(W_\lambda\left(BN_s\left(F_2\right)\right)\right), W_\lambda = \lambda_i / \sum_{j=0} \lambda_j \tag{4}$$

The global scheduling mechanism of NAM improves the performance of deep convolutional networks by reducing information reduction and amplifying the global interaction representation, reducing the weight of non-salient features more efficiently and improving the accuracy of foreign body detection. Given the sparse weight penalty applied to the attention module, the weight calculation is more efficient whilst maintaining the same feature enhancement performance.

2) MRB MODULE

MRB can fully extract the feature information of the image with a small number of parameters. Its specific structure is shown in Figure 4, which is mainly composed of three submodules, namely, the local information encoding module, the global information encoding module and the feature fusion module.

The local information encoding module initially models the local features of the input feature map X of size $H\times W\times C_{in}$ through a 3×3 depth convolution layer, and then uses a convolution layer with a convolution kernel size of 1×1 to adjust the number of channels. The global information encoding module performs global feature modelling by an Unfold-Transformer Block-Fold structure, after which the number of channels is adjusted back to the original size by a convolutional layer of size 1×1. To reduce the number of parameters, we use the dot product operation of pixels in was applied to the same position in the Transformer block to avoid information redundancy caused by the self-attention operation on all pixels and further improve the detection speed of model, which is shown in Figure 5.

The feature fusion module performs concatenation(Concat) splicing on the output feature maps after global and local feature extraction through a shortcut branch, and then performs residual connection with the original input feature map after dimension reconstruction to obtain the final output graph Y with the size of $H\times W\times C_{out}$. The output after the fusion of local features and global features is called an intermediate fusion feature. The reason why it is fused with the input feature residual is to ensure that the input feature is independent of the local and global features at other positions in the feature map, so as to simplify the learning task of the fusion module and enhance the training efficiency of the model. In addition, experiments show that the introduction of residual connections in the new MRB architecture is helpful for improving the detection accuracy.

B. FEATURE FUSION MODEL

The traditional feature pyramid network (FPN) [28] can greatly improve the detection performance by using simple up-sampling, down-sampling and directional splicing of feature channels. However, for a lightweight detection network with less feature information, we need to design richer feature linking methods and high-performance feature processing modules to improve the extraction efficiency of high-dimensional semantic information and enhance the ability of the model to detect and identify intraocular foreign bodies.

In this study, the PAFPN based on FPN is used to fuse multiscale features, which makes it easier to transfer the bottom information to the high level by introducing the bottom-up path to realize the efficient fusion of different levels of features. To this end, the spatial pyramid pooling fusion cross-stage partial connection (SPPFCSPC) module is optimized by reconstructing the spatial pooling layer and the convolutional layer connections for reducing the computational overhead. And the multiple concatenation(Multi-concat) module is used for fusing diverse features and further enrich feature information. In particular, two lightweight modules, namely, the spatial-to-depth convolution(SPD-Conv) and convolutional mixing module(ConvMix) are introduced to capture the key semantic features in spatial and channel directions. Some important modules in the PAFPN are described below.

1) MULTI-CONCAT AND SPPFCSPC MODULES

Figure 6 shows the specific structure of the Multi-Concat and SPPFCSPC modules, where the CBS submodule is a fundamental convolution calculation module consisting of a Conv layer, a BN layer and a SiLU activation function. The Multi-Concat structure is composed of multiple CBS submodules, which improve the learning ability of the network by guiding the calculation blocks of different feature groups to learn more diverse features while preserving the original gradient path. In the Multi-Concat module, the input of the final stacking part contains multiple branches. The lower branch is a CBS module with a convolution kernel size of 1 that compress the number of feature channels to half, while the upper branch employs the same CBS module initially for channels adjustment and subsequently connects four CBS modules with a convolution kernel size of 3 to extract feature information. The four CBS modules reduce the number of feature channels to a quarter and output them to the final stacked section following the hierarchy. Once the above output feature layers are stacked, they are reintegrated by a standardized convolution to obtain the final output. The denser residual structure of multi-branch stacking corresponds to an improved accuracy by increasing the depth. At the same time, the various residual blocks are constructed in a stacked pattern of skip connections, effectively mitigating

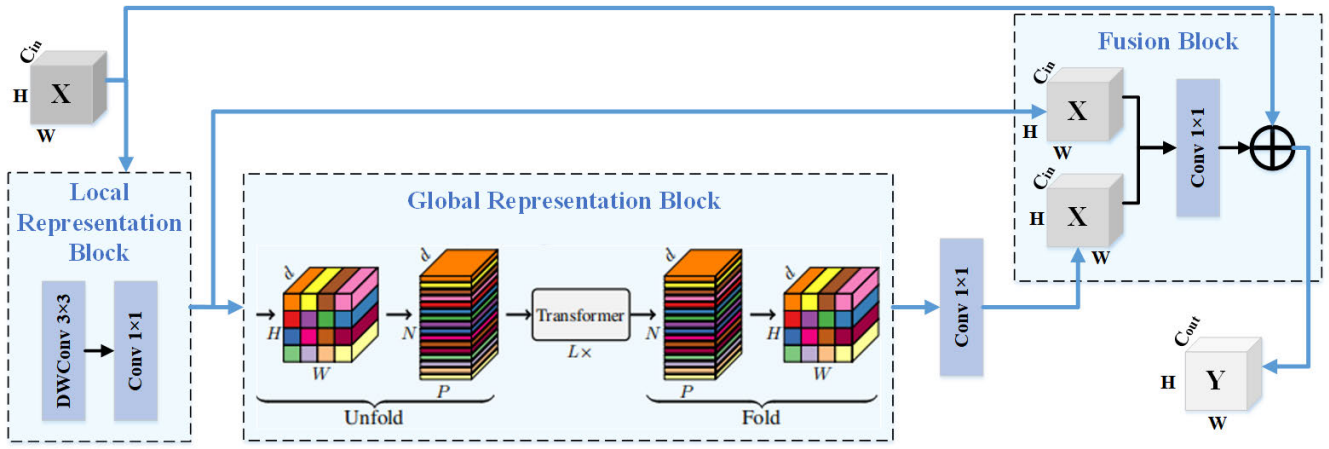

FIGURE 4. MRB module.

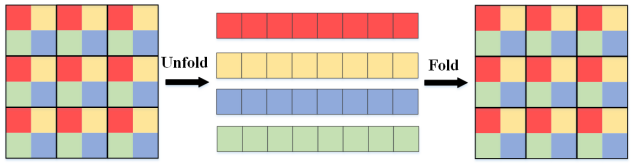

FIGURE 5. Same position pixel dot product operation.

the issue of gradient vanishing that arises with increased depth.

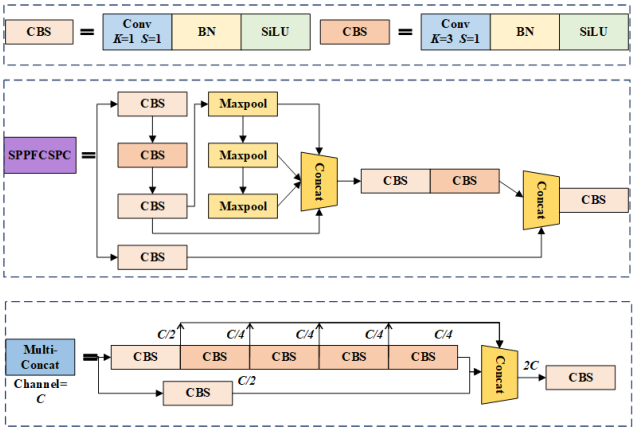

FIGURE 6. Multi-Concat and SPPFCSPC modules.

Unlike the typical spatial pyramid pooling structure, the SPPFCSPC module is redesigned to expand the receptive field. Meanwhile, the pooling window size and the connection mode of Maxpool are adjusted, which have larger residual edges for auxiliary optimization and feature extraction and further reduce the computational overhead. Initially, the input feature is divided into two branches: the lower branch undergoes regular convolution, and the upper branch utilizes spatial pyramid pooling. Finally, the two branches are merged and connected, which can improve the detection accuracy whilst halving the amount of calculation. Specifically, the SPPFCSPC module incorporates multiple Maxpool operations into a sequence of CBS convolutions. It rearranges the three Maxpool layers that were arranged in a parallel structure into serial connections, and uniformly sets the size of the pooling window to 5. This enables better detection of small foreign bodies, while also avoiding image distortion.

2) SPD-CONV MODULE

The performance of the traditional feature pyramid structure composed of a common CNN rapidly degrades when tackling small object detection tasks in complex environments due to the loss of fine-grained information caused by strided convolution and pooling operations, and the low efficiency of feature representation learning. The SPD-Conv module includes a Space-to-depth(SPD) layer and a Non-strided Convolution(Nconv) layer, thereby obviating the convolution step and the pooling operation. Incorporation the SPD-Conv into PAFPN can enhance the network's ability to capture small foreign bodies.

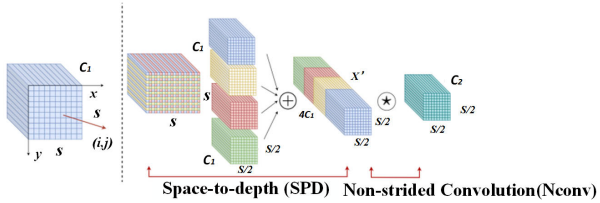

FIGURE 7. Schematic diagram of the SPD convolution operation.

Figure 7 illustrates how SPD-Conv cooperates with a down-sampling factor scale of 2.Fristly, the SPD layer splits the input features of $S\times S\times C_1$ into four subfeature maps.Then,it transforms these maps into the reconstructed features of ${}^{s}/_{2}\times{}^{s}/_{2}\times 4C_1$.The Nconv layer continues to further convert the reconstructed feature map X' into the

output feature of ${}^{s}/_{2} \times {}^{s}/_{2} \times C_2$ by utilizing nonstrided convolution with a stride of 1, where $C_2 < 2^2 \times C_1$.The Nconv can retain all discriminative feature information to the greatest extent and prevent nondiscriminatory loss of information. Substituting convolution step and pooling layers with SPD-Conv can effectively improve the detection accuracy while also reducing computational parameters. In addition, the network fused with SPD-Conv exhibits better generalisation, which reduces the workload of model hyperparameter tuning.

3) CONVMIX MODULE

High-dimensional features in deep networks tend to possess more feature channels and contain abundant semantic feature information. Considering that the repetition of gradient information during network optimisation leads to an increase in the computational overhead of inference, we suggest the application of the ConvMix module to optimize the gradient information transmission while lowering the number of inference calculations. The ConvMix employs a dual-branch structure, as shown in Figure 8. The upper branch compresses the feature channel through the CBS unit, reducing the amount of calculation before transmitting it to the DWConv processing unit. The lower branch also uses the CBS unit to compress the number of channels to one-half and then performs Concat splicing with the output feature map of the upper branch so that the model can learn enough feature information. The final connected CBS unit is used to adjust the final output size to be consistent with the input.

The ConvMix module initially divides the feature maps of the base layer into two sections and then merges them by crossing the stage hierarchy to decrease the amount of computation whilst ensuring accuracy. The input is connected to the DWConv result by a residual structure and then by a pointwise convolution with a convolution kernel of 1. After two convolution operations, the GELU activation function is connected with the BN layer. The ConvMix module designed on the basis of DWConv can greatly reduce the calculation overhead of the model and realize the lightweight reconstruction of the PAFPN.

The final output shown in Figure 1 is the classifier and regressor for obtaining the final location and type information of intraocular foreign bodies. Three enhanced effective feature layers can be fused by PAFPN. At this time, the feature map can be regarded as a set of feature points, each feature point has a target prior box and each prior box has several features of the channel. The classifier and regressor can determine whether the prior box at the feature point corresponds to an object, which is implemented in a 1×1 convolution.

III. RESULTS AND DISCUSSION

A. DATASET

Based on the study and analysis of 82 patients' brain CT [29] images, we selected 931 typical CT images of intraocular foreign bodies to generate the initial dataset. Among them, 652 lesions were related to foreign bodies, 400 lesions to air accumulation, and 121 lesions to blood accumulation. Some images contained more than one type of intraocular foreign body, as shown in Figure 9.

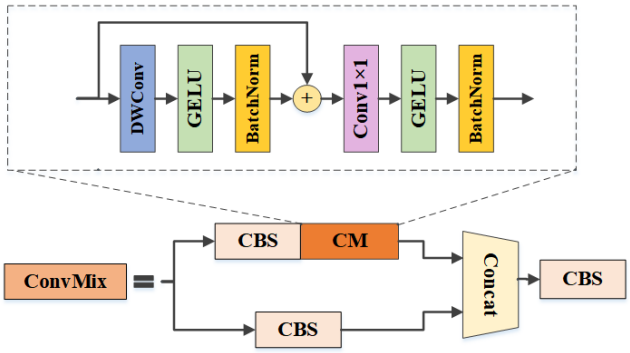

FIGURE 8. ConvMix module.

Given the specificity of medical images, the CT image dataset of intraocular foreign bodies manifests the problems of a small sample size and an imbalanced number of positive and negative samples. The performance of the object detection network based on DL shows a positive correlation with the number of images present in the dataset. Expanding the original dataset through the use of data augmentation technology can effectively improve the performance of the network in detecting and recognizing intraocular foreign bodies, even in complex environments.

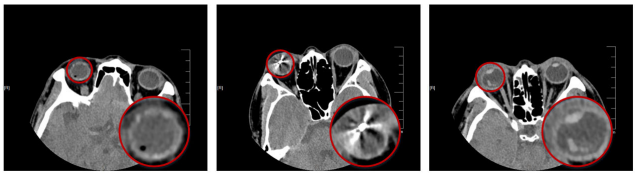

FIGURE 9. Different types of intraocular foreign bodies in brain CT images.

As depicted in Figure 10, the intraocular foreign body only appears in the inside and edge area of the patient's eyeball, which corresponds to the upper eyeball part in the whole CT image, that is, the key area where the target to be detected appears. Therefore, by maximizing the utilization of the key detection area of the image, the imbalance between the number of positive and negative samples in the dataset can be effectively improved, ensuring that the target detection network extracts sufficient lesion features. In this study, we cross-segmented each CT image in the original intraocular foreign body dataset to obtain four image slices of the same size. The scattered image slices were reorganized into the original size by randomly splicing them together, both in key areas and throughout the cross-cut image slice set. Only the CT image slice containing the eyeball region was randomly recombined by slice random stitching. The reconstructed

image sizes of from both methods were consistent with the original image. The above scheme is an effective solution for addressing the difficulty of training with limited samples.

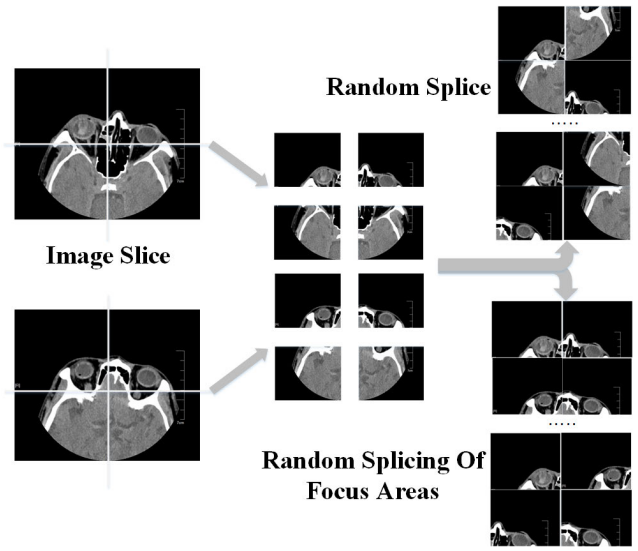

FIGURE 10. Data augmentation of CT images of intraocular foreign bodies.

By segmenting and reconstructing CT images randomly, a larger dataset can be generated, thus increasing the network model's resilience. Furthermore, for the particular dataset relating to intraocular foreign bodies, the randomized slicing approach in pivotal areas significantly boosts target utilization efficiency, thereby augmenting feature extraction, detection, and recognition abilities of the model within intricate environments. Data augmentation was performed on the original dataset, expanding it to 10,587 images. The training set and the testing set are divided at a ratio of 8:2. In addition, we selected 500 CT images that did not contain intraocular foreign bodies as the specificity test data set to calculate the misjudgment ratio of the proposed model.

B. MODEL TRAINING AND CONVERGENCE ANALYSIS

To decrease the model training time, the pretrained parameter values on the COCO public dataset are used as the initial values for the proposed model parameter training. This model has an input image size of 640×640 and a network layer consisting of 448 layers with Adam as the optimizer. The initial learning rate for the model is set to 0.001, the batch size is 32 and the training epoch is 300.

The model training process mainly contains three aspects of loss, namely, rectangular loss (loss-box), confidence loss (loss-object) and classification loss (loss-class). The rectangle represents the size and location of the object, whilst the confidence measure displays the level of confidence in the predicted rectangle, evaluated within a range of 0 to 1. Higher values indicate a greater probability of finding the target within the rectangle. The classification probability defines the class of the object. The network loss is a combination of three losses that are weighted. Firstly, we calculate the

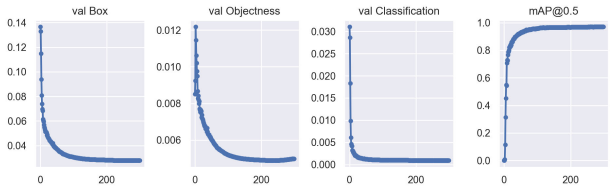

FIGURE 11. Loss function and mAP@0.5 index changes during model training.

rectangular box loss by using the complete intersection over union(CIOU) loss. Secondly, we use the binary cross entropy(BCE) loss to calculate the confidence loss and classification loss. As indicated in Figure 11, the rectangular box loss and classification loss decrease significantly with increased training rounds, while the confidence loss curve exhibits greater fluctuations during the initial stage of training and later decreases smoothly. The mAP@0.5 indicator curve rises smoothly with training rounds, reaches a maximum at approximately 150 rounds and then remains flat.

C. ABLATION EXPERIMENTS

In this study, we propose a real-time method for detecting and recognizing intraocular foreign bodies, which employs MobileViTv3 network as the feature extraction model and PAFPN as the feature fusion model. To more clearly understand the influence of each module of the models on the overall detection and recognition effect, we conducted ablation experiments and summarized results are shown in Table 1.

TABLE 1. Comparison of ablation study performance.

Method	Para.(M)	mAP@0.5	Detection times (ms/sheet)
MobileViTv3	15.88	84.4	3.1
MobileViTv3+IRB	15.88	84.9	3.2
MobileViTv3+PAFPN	35.09	97.0	5.7
MobileViTv3+IRB+PAFPN	35.24	96.8	5.8
MobileViTv3+IRB+PAFPN +SPD	33.19	97.1	5.2
MobileViTv3+IRB+PAFPN+ ConvMix	32.6	96.6	5.2
MobileViTv3+IRB+PAFPN+ SPD+ConvMix(Ours)	**30.96**	**97.2**	**5.0**

The combination of the IRB structure and NAM module improves the mAP@0.5 index by 0.5 without intensifying the computational burden of the network. Additionally, the detection time of the model remain at a lightweight level. Thereafter, on the foundation of the backbone network, the PAFPN structure is added for feature fusion between different scales, and the detection performance of the model is greatly improved. The mAP@0.5 is increased to 97.0, but it inevitably causes a surge in the number of model parameters, and the detection time of the model is greatly reduced. The detection time per CT image increases from 3.1 to 5.7 ms/sheet. Therefore, we introduce the SPD module

TABLE 2. Performance comparison of models using different feature extraction networks.

Backbone	mAP@0.5	Precision	Sensitivity	Specificity	Para.(M)	Detection times (ms/sheet)
MobileNetv1 [19]	84.3	91.7	84.6	62.4	**20.50**	3.9
MobileNetv2 [20]	88.6	86.7	89.9	66.0	21.51	4.3
MobileNetv3 [18]	87.0	**94.7**	91.0	74.8	21.05	3.8
ShuffleNetv2 [24]	61.4	78.3	62.3	87.4	33.80	**0.8**
GhostNetv2 [23]	68.7	77.8	67.8	82.8	60.11	1.6
MobileViTv1 [22]	93.6	91.6	98.0	81.8	26.67	4.8
MobileViTv2 [32]	92.3	88.1	93.4	78.2	22.09	4.5
MobileViTv3 [21]	96.7	92.2	97.0	82.4	31.94	5.2
Ours	**97.2**	93.5	**98.0**	**88.0**	30.96	5.0

and the ConvMix module to lightly reshape the feature fusion structure to improve the detection speed. The SPD convolution structure optimizes the down-sampling procedure in the deep network to avoid the enormous computational overhead caused by high-dimensional features. When the SPD convolution structure is added to PAFPN, the model parameters are reduced by 2 M and the detection time is reduced by 0.6 ms/sheet.

The ConvMix module enhances the transmission of gradient information in deep network, reducing the amount of inference calculation. Integrating the ConvMix module into PAFPN reduces the model parameters by 3 M and the inference time by 0.6 ms/sheet, but a slight decrease in mAP@0.5 is also observed. It is possible that the ConvMix module possesses a weaker capability to extract feature information at the lower part of the network when compared to both the feature extraction module in YOLOv5l [30] and feature extraction module in YOLOv7l [31]. However, the SPD module can address this issue due to its competence in exploring the feature depth information. Therefore, incorporating the SPD module and ConvMix module into the PAFPN structure leads to improved detection performance and reduced detection time. Our model achieves an mAP@0.5 of 97.2 and a detection time of 5.0 ms/sheet, which is 0.7 less than that of the traditional PAFPN. This advancement is beneficial in promoting clinical injury detection. In addition, as shown in Fig. 12, we compared the actual visual detection effect graphs of MobileViTv3+IRB and MobileViTv3+IRB+PAFPN for intraocular foreign bodies with our model, and ours exhibits superior foreign object position detection accuracy. Therefore, the combination of this model is optimal in terms of the number of parameters, mAP@0.5, detection time and location accuracy.

D. COMPARISON OF ALGORITHMS

1) PERFORMANCE COMPARISON OF MODELS BY USING DIFFERENT FEATURE EXTRACTION NETWORKS

To evaluate the effectiveness of the MobileViTv3 as a feature extraction network in detecting and recognizing intraocular foreign bodies, this study compares it with eight other lightweight models, including MobileNetv1 [19], MobileNetv2 [20], MobileNetv3 [18], ShuffleNetv2 [24], GhostNetv2 [23], MobileViTv1 [22], MobileViTv2 [32], and MobileViTv3 [21]. The comparison is carried out in this section.

Table 2 illustrate the superior performance of our proposed model in terms of mAP@0.5, sensitivity and specificity. Meanwhile, the accuracy is only 1.2 lower than that of the original MobileNetv3 used as the feature extraction network. Our analysis of MobileNetv3's low sensitivity and specificity indexes has revealed that its poor recall ability yields a high accuracy rate for foreign bodies detected in a limited range, ultimately leading to a falsely high detection accuracy. In regards to accuracy and sensitivity for clinical examination, the MobileNet series, GhostNetv2, and ShuffleNetv2 fail to meet the necessary requirements. Our model, compared to the original MobileViTv3, reduces the number of parameters by approximately 1 M and the detection time by 0.2 ms/sheet. Meanwhile, compared with MobileViTv1 and MobileViTv2, the difference in detection time of the proposed model is less than 0.5 ms/sheet, while showing a significant advantage in mAP@0.5.

2) PERFORMANCE COMPARISON OF MODELS USING DIFFERENT FEATURE FUSION NETWORKS

In order to verify the effectiveness of the enhanced PAFPN developed in this study, we have chosen to carry out comparative experiments using four different feature fusion networks, namely, PANet+YOLOv5l [33], PANet+YOLOv5l, ASFF [34] and BiFPN [35]. In this section, we only concentrate the performance of the feature fusion network, thus adopting the improved MobileViTv3 is selected as the backbone for all models.

TABLE 3. Performance comparison of models using different feature fusion networks.

FPN	mAP@0.5	Precision	Sensitivity	Specificity	Para. (M)	Detection times (ms/sheet)
PANet+YOLOv5l	96.6	90.2	97.9	83.6	34.83	5.7
PANet+YOLOv7l	97.0	93.1	98.0	86.4	39.15	7.0
ASFF [34]	92.6	93.2	86.3	77.2	42.86	7.2
BiFPN [35]	97.2	91.3	97.8	86.4	44.42	7.6
Ours	**97.2**	**93.5**	**98.0**	**88.0**	**30.96**	**5.0**

According to Table 3, the feature fusion network proposed achieves superior results in all indicators, with significant

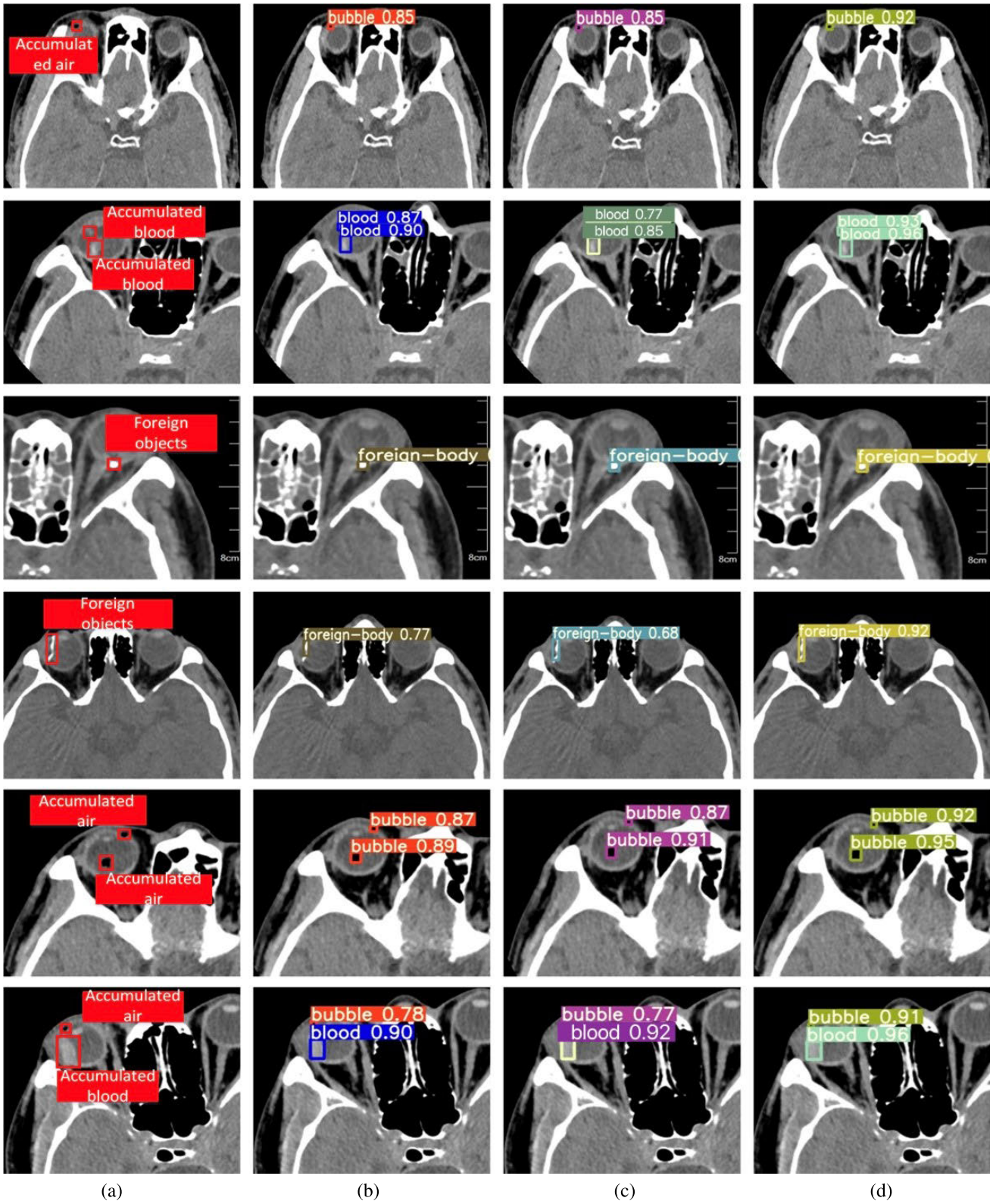

FIGURE 12. Comparison of visual effects of ablation experiment (a)Ground Truth (b)MobileViTv3+IRB (c)MobileViTv3+IRB+PAFPN (d)Ours.

advantages in its detection performance and speed. Notably, the BiFPN model performs similarly to this model in terms of mAP@0.5, but its high large number of parameters and long detection time render it unsuitable for real-time

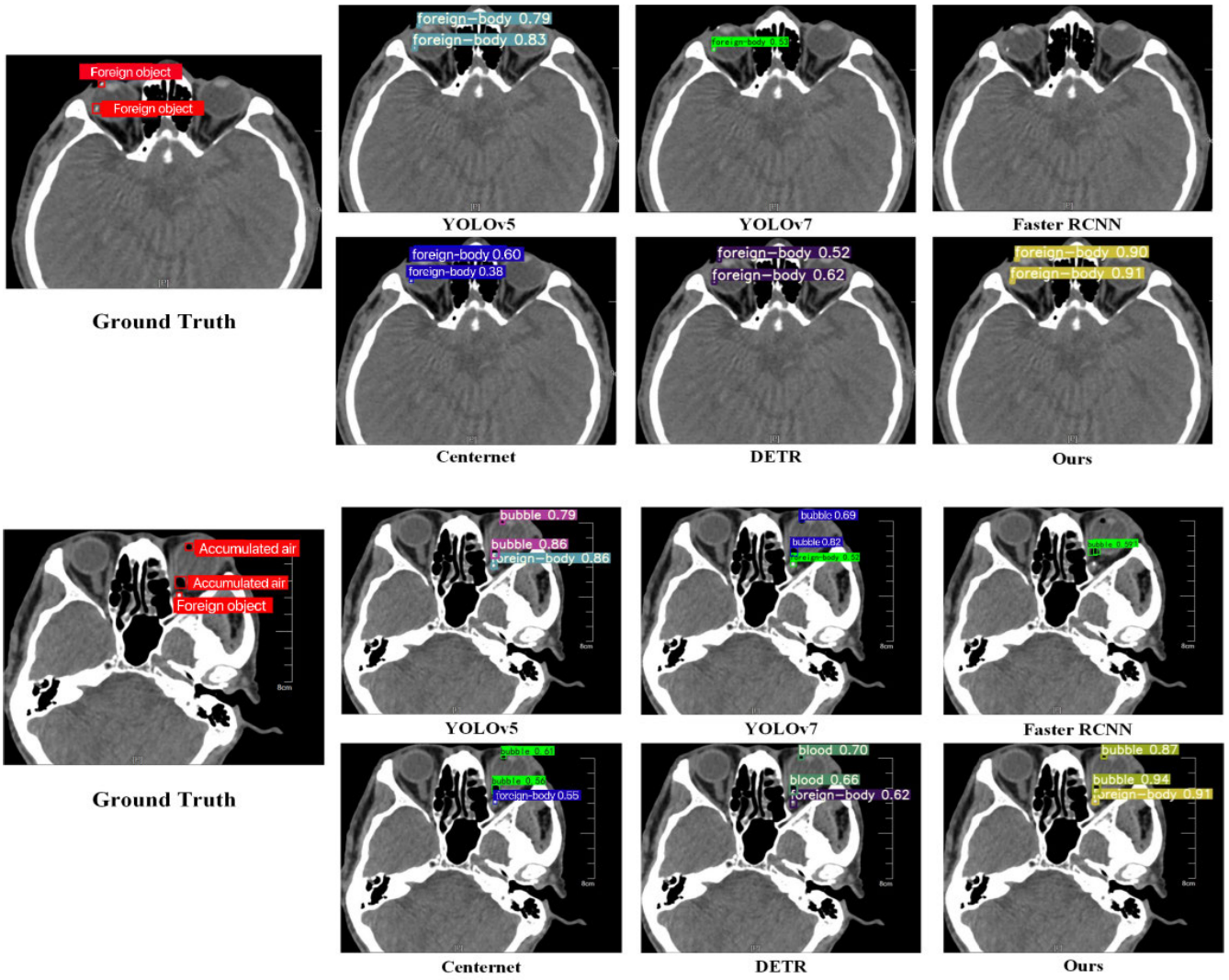

FIGURE 13. Comparison of visualization effects of different object detection networks.

clinical detection. The combined models of PANet with two YOLO feature fusion networks also display good detection efficacy. In comparison to the BiFPN structure, which has a significant computational overhead, this study finds the parameter number and detection time of the proposed method to be acceptable. Furthermore, in the desighn of the feature fusion network, this study introduces two high-performance convolution processing modules, Multi-Concat and SPPCSPC, to lightly reshape the network structure while ensuring the detection effect. As a result, the proposed model achieves faster speed and better detection and recognition performance.

3) PERFORMANCE COMPARISON WITH OTHER OBJECT DETECTION ALGORITHMS

In this section, we compare the performance indicators of various object detection networks in the detection of intraocular foreign bodies to validate the effectiveness of our model. Figure 13 displays the visual effects of the algorithm comparison. The comparison algorithms are primarily divided into three categories. The study evaluated five single-stage object detection algorithms namely, YOLOv5l [30], YOLOv7l [31], YOLOX [26], CenterNet [36] and CornerNet [37]; a two-stage object detection approach, Faster RCNN [38]; and a fully transformer-based object detection algorithm, DETR [39]. Table 4 shows that our model achieved the best results in various detection performance indicators. In particular, the proposed model is faster in terms of detection time than the single-stage detection networks, and more accurate in terms of detection accuracy than the two-stage detection network. Furthermore, our model has obvious advantages in the specificity index, indicating that its false positive target false detection problem has also been improved.

TABLE 4. Performance comparison of different object detection networks.

Network	mAP@0.5	Precision	Sensitivity	Specificity	Detection times (ms/sheet)
YOLOv5l [30]	96.5	**97.9**	93.8	84.0	7.6
YOLOv7l [31]	83.9	92.3	59.4	82.2	10.9
YOLOX [26]	90.4	93.1	87.5	83.4	8.3
Faster RCNN [38]	83.9	89.3	88.6	64.0	43.1
SSD [40]	62.1	78.9	52.1	87.6	17.8
CenterNet [36]	68.8	90.5	64.8	65.4	7.8
CornerNet [37]	70.1	91.3	64.7	66.0	7.9
DETR [39]	81.3	88.7	76.5	81.2	11.2
Ours	**97.2**	93.5	**98.0**	**88.0**	**5.0**

IV. SUMMARY

To address the issues of imprecise marking, large positioning error, cumbersome operation process and high detection time during the detection of intraocular foreign bodies in CT images by current technical means, a lightweight detection and recognition model based on feature exaction and fusion is proposed in this study. We use MobileViTv3 as the backbone network and utilize the PAFPN to fuse multiscale feature information. By redesigning and incorporating multiple modules, a lightweight model for real-time detection and recognition is constructed, which realizes the efficient detection and diagnosis of intraocular foreign bodies injuries in brain CT images.

The experiments indicate that the proposed model developed in this research has a mere 30 M parameters and detection time of 5.0 ms/sheet. Compared with other lightweight models, it has lower computational cost and faster detection speed, which allows the proposed model to run in real time on poorly configured hardware. Furthermore, in comparison to various object detection algorithms, this model achieves the highest mAP@0.5 of 97.2, sensitivity of 98.0 and accuracy of 93.5. The specificity index of 88 also demonstrates that this model effectively addresses the issue of high false positive rate of detection results. This proposed model presents significant advantages in the detection and recognition of intraocular foreign bodies, making it well-suited for clinical applications.

REFERENCES

[1] L. Ma, G. Xiu, J. Muscat, R. Sinha, D. Sun, and G. Xiu, "Comparative proteomic analysis of exhaled breath condensate between lung adenocarcinoma and CT-detected benign pulmonary nodule patients," *Cancer Biomarkers*, vol. 34, no. 2, pp. 163–174, May 2022.

[2] H. C. Jung, S. Y. Lee, C. K. Yoon, U. C. Park, J. W. Heo, and E. K. Lee, "Intraocular foreign body: Diagnostic protocols and treatment strategies in ocular trauma patients," *J. Clin. Med.*, vol. 10, no. 9, p. 1861, Apr. 2021.

[3] X. Liu, Q. Bai, and X. Song, "Clinical and imaging characteristics, outcomes and prognostic factors of intraocular foreign bodies extracted by vitrectomy," *Sci. Rep.*, vol. 13, no. 1, Aug. 2023, Art. no. 14136.

[4] M. I. N. M. Hilal, R. Ganesan, N. M. Norsuddin, M. I. Ibrahim, S. M. S. S. Rahmat, M. K. A. Karim, and I. N. C. Isa, "Radiation dose to the eye and potential occurrence radiation-induced cataract following computed tomography (CT) head examination," *Malaysian J. Public Health Med.*, vol. 21, no. 2, pp. 1–7, 2021.

[5] Y. Liu, S. Wang, Y. Li, Q. Gong, G. Su, and J. Zhao, "Intraocular foreign bodies: Clinical characteristics and prognostic factors influencing visual outcome and globe survival in 373 eyes," *J. Ophthalmol.*, vol. 2019, pp. 1–7, Feb. 2019.

[6] F. R. Imrie, A. Cox, B. Foot, and C. J. MacEwen, "Surveillance of intraocular foreign bodies in the U.K.," *Eye*, vol. 22, no. 9, pp. 1141–1147, Sep. 2008.

[7] 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.

[8] W. Lin, J. Chu, L. Leng, J. Miao, and L. Wang, "Feature disentanglement in one-stage object detection," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109878.

[9] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proc. IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[10] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[11] R. Zhang, D. Tian, D. Xu, W. Qian, and Y. Yao, "A survey of wound image analysis using deep learning: Classification, detection, and segmentation," *IEEE Access*, vol. 10, pp. 79502–79515, 2022.

[12] A. Khan, H. Kim, and L. Chua, "PMED-net: Pyramid based multi-scale encoder–decoder network for medical image segmentation," *IEEE Access*, vol. 9, pp. 55988–55998, 2021.

[13] X. Zhao, X. Shen, W. Wan, Y. Lu, S. Hu, R. Xiao, X. Du, and J. Li, "Automatic thyroid ultrasound image classification using feature fusion network," *IEEE Access*, vol. 10, pp. 27917–27924, 2022.

[14] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, "Artificial intelligence and deep learning in ophthalmology," *Brit. J. Ophthalmol.*, vol. 103, no. 2, pp. 167–175, Feb. 2019.

[15] X. Pan, K. Jin, J. Cao, Z. Liu, J. Wu, K. You, Y. Lu, Y. Xu, Z. Su, J. Jiang, K. Yao, and J. Ye, "Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning," *Graefe's Arch. Clin. Exp. Ophthalmol.*, vol. 258, no. 4, pp. 779–785, Apr. 2020.

[16] L. Lin, Z. Wang, J. Wu, Y. Huang, J. Lyu, P. Cheng, J. Wu, and X. Tang, "BSDA-Net: A boundary shape and distance aware joint learning framework for segmenting and classifying OCTA images," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Strasbourg, France, Springer, Sep./Oct. 2021, pp. 65–75.

[17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.

[18] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

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

[21] S. N. Wadekar and A. Chaurasia, "MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," 2022, *arXiv:2209.15159*.

[22] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.

[23] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetv2: Enhance cheap operation with long-range attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 9969–9982.

[24] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

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

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

[27] 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*.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[29] N. Shi-Huai, Y. Xiao-Ping, and C. Q. Yu, "Evaluation of X-ray, B-ultrasound and CT for localization and diagnosis of foreign bodies in the eye wall," *China Practical Med.*, vol. 8, no. 23, pp. 100–101, 2013, doi: 10.3969/j.issn.1673-7555.2013.23.069.
[30] 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.
[31] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.
[32] S. Mehta and M. Rastegari, "Separable self-attention for mobile vision transformers," 2022, *arXiv:2206.02680*.
[33] 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.
[34] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.
[35] 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. 10781–10790.
[36] 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. 6569–6578.
[37] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.
[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 213–229.
[40] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Springer, Oct. 2016, pp. 21–37.

YIRAN LIU is currently pursuing the M.M. degree with Nanjing Medical University, Nanjing, Jiangsu, China. Her research interest includes biomedical image processing. She has won the second prize and third prize in Jiangsu Translation Competition.

YITING ZHENG is currently pursuing the M.S. degree with the China University of Mining and Technology, Xuzhou, Jiangsu, China. Her research interest includes image processing.

XIAOYU ZHU is currently pursuing the M.S. degree with the China University of Mining and Technology, Xuzhou, Jiangsu, China. Her research interest includes medical image processing.

JUNZHE CHEN received the master's degree from the China University of Mining and Technology, Xuzhou, Jiangsu, China. His research interest includes image processing.

SUYAN LI is currently a Professor with Xuzhou No. 1 People's Hospital. She is also a Key Medical Talent in Xuzhou. Her research interest includes ophthalmology image processing.

ZHAOLIN LU received the M.S. degree from Xidian University, in 2006, and the Ph.D. degree from the China University of Mining and Technology, Xuzhou, Jiangsu, China.

He is currently an Associate Professor of Xuzhou No.1 people's Hospital (Affiliated Hospital of China University of Mining and Technology). He has published over 20 papers in domestic and international academic journals and conference proceedings. His research interest includes image processing.