IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 May 2023, accepted 31 May 2023, date of publication 14 June 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3286031

RESEARCH ARTICLE

MBAB-YOLO: A Modified Lightweight Architecture for Real-Time Small Target Detection

JUN ZHANG, YIZHEN MENG, XIAOHUI YU, HONGJING BI, ZHIPENG CHEN, HUAFENG LI, RUNTAO YANG, AND JINGJUN TIAN

Computer Science Department, Tangshan Normal University, Tangshan 063000, China

Corresponding author: Yizhen Meng (xingowenanq@aliyun.com)

This work was supported in part by the Science and Technology Program of Tangshan under Grant 22130214G, in part by the Scientific Research Foundation of Tangshan Normal University under Grant 2022C47 and Grant 2023C15, in part by the Science and Technology Plan Project of Tangshan Science and Technology Bureau under Grant 21130212D, in part by the Ph.D. Foundation of Tangshan Normal University under Grant 2023B06, and in part by the Research Project of Education and Teaching Reform of Tangshan Normal University under Grant 2022JG14 and Grant 2022JG10.

ABSTRACT Current target detection methods have achieved high accuracy for detecting large and medium-sized targets. However, due to factors such as the small number of pixels and features available for targets in images, the detection performance for small targets is generally unsatisfactory. In addition, the real-time performance of target detection is also critical. In conclusion, a modified lightweight architecture for real-time small target detection, i.e., MBAB-YOLO, is proposed based on You Only Look Once (YOLO) model by combining channel-wise attention block, space-attention block and multi-branch-ConvNet (Convolutional Neural network) structure. Specifically, our method is more suitable for the rich scale information of small targets through proposed adaptive multi-receptive-field focusing, and then combines proposed blended attention block (BAB) to re-calibrate small target information to make it more prominent and improve the discriminability of small target features. Finally, extensive experiments have been conducted on the open source data set for the proposed real-time small target detection method, i.e., MBAB-YOLO. The results of ablation experiment and contrast experiment show that our method has excellent performance, not only with high detection accuracy, but also with fast detection speed. Compared with the various benchmark methods, it achieves a good trade-off between the two aspects mentioned above. In addition, this paper gives a comprehensive and detailed review of the current work about small target detection from different several perspectives, which can be used as a reference for future researchers.

INDEX TERMS Deep learning, target detection, channel-wise attention, space-attention, YOLO.

I. INTRODUCTION

With the continuous development of deep learning and the constant reduction of hardware cost, deep learning-based target detection methods have made significant progress. Compared to medium and large target detection, small target detection has the characteristics of less target feature information, imbalanced data distribution, and susceptibility to environment, which lead to low accuracy in small target detection. Small target detection has extensive applications in tasks such as maritime rescue, surveillance recognition, unmanned aerial vehicle (UAV) identification, remote sensing satellite, and marine life detection. Therefore, studying the small target detection method and improving its accuracy and efficiency is of great significance.

Due to the successive down-sampling operation, deep learning-based target detection method filters the correlated noise during feature extraction, enhancing the feature representation of the target, while also causing small targets to lose information in the forward propagation of the network. To this end, some scholars have proposed different multi-scale feature fusion structures based on feature pyramid network

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

(FPN) [1], such as path aggregation network (PANet) [2], neural architecture search network (NAS-Net) [3], deep feature pyramid networks (DFPN) [4], Bidirectional Feature Pyramid Networks (BiFPN) [5], etc. However, among these networks, the fusion between different layers is only simple summation, ignoring the relevance of the target in the scene, which has limited improvement for small target detection. Specifically, squeeze excitation network (SE-Net) [6], convolutional block attention module (CBAM) [7], frequency channel attention network (FcaNet) [8] and other methods model small targets from different perspectives of channel-wise attention and space-attention to obtain attention weight matrices in two dimensions, thus enhancing small target feature representation and suppressing other targets and complex environmental information. However, these attention network designs ignore the effect of different convolutional kernels on small target detection.

To address the above problems and enhance the real-time performance of the network, this paper proposes a lightweight architecture for real-time small target detection, i.e., MBAB-YOLO. Specifically, the method uses YOLOv5s (the s version of You Only Look Once model) as the baseline structure for small target detection, and then improves it with proposed blended attention block (BAB) and multi-branch-ConvNet (Convolutional Neural network) structure. The main contributions of this paper are as follows:

1) We combined channel-wise attention (CA) block and space-attention (SA) block, and reorganized the connection structure to propose BAB. BAB can obtain the rich global spatial attention weight matrix, enhance small target feature information, and suppress irrelevant information such as the background.

2) We proposed a novel multi-branching blended attention block (MBAB) by combining multi-branch-ConvNet structure and BAB mechanism. MBAB can adaptively adjust the receptive field size based on the scale of the input target, and enhance the feature representation of small targets.

3) To improve the feature extraction capability for small targets, we improved the core residual block, i.e., C3, of YOLOv5 and combined MBAB with C3 to propose a feature extraction residual block based on CSPNet (cross stage partial network) (CSP-MBAB, abbr, CMBAB). CMBAB can focus more attention on small targets during feature extraction, enhancing the feature information of small targets. Meanwhile, a new prediction branch and small target detection head are introduced in the P2 layer, which has more shallow information and is beneficial for small target detection.

4) This paper comprehensively introduced the general and the specific research status of small target detection, as well as YOLO family, which can undoubtedly serve as a significant reference for later researchers.

5) Extensive contrast experiments and ablation experiments have verified the trade-off between accuracy and efficiency of proposed method, demonstrating its superiority as a lightweight architecture for real-time small target detection.

II. RELATED WORKS

The small target in the COCO dataset is an absolute definition, which refers to the target that smaller than 32×32 pixels in an image. However, in practical applications, a more common approach is to use the ratio of the target size to the original image, which is referred to as a relative definition. For example, the target with a ratio smaller than 0.1 can be considered the small. In general, there is no strict definition for the small target, and it needs to be determined based on the actual engineering application.

In the development of target detection, it has been gradually discovered that detecting the small target is more challenging than detecting the medium to large target.

Regardless of whether it is a relative or absolute definition, the small target typically has fewer pixels, lower resolution, and lack feature information. After continuous exploration, several reasons that contribute to the low detection accuracy of small target have been revealed:

(1) Lack of feature information: Due to the small number of pixels in the small target, deep neural networks, which undergo dozens or hundreds of convolution and pooling operations, will down-sample the image to reduce the computational cost and expand the receptive field, generating the thumbnail image. However, this will cause a significant loss of information in the small target, making the information of the small target in the feature map less and less.

(2) Information loss in forward propagation of neural networks: During the forward propagation procedure, the semantic information of the feature map becomes stronger while the positional information gradually gets lost, making it difficult to locate the coordinates of the target.

(3) Unequal distribution of sample quantities in the dataset: If the number of small target in the training set is distributed unevenly compared to the medium and large target, it will result in the network having lower adaptability to different sizes of the target during learning, leading to a decrease in detection accuracy. In the COCO dataset, images containing targets of all three sizes (small, medium, and large) account for 52.3% of the total samples, with the proportion of large & medium targets and small targets being 70.7%, 83.0%, respectively. This reasonable distribution of samples makes the COCO a common dataset for small target detection. Obtaining an class-balanced dataset is also a major challenge in practical applications.

(4) Setting of anchors: Due to the varying sizes and aspect ratios of targets to be detected, it is difficult to set anchors that match the actual situation. Existing methods use multiple sets of anchors or calculate anchors based on the training data set, but the generalization ability is poor when detecting unseen targets.

(5) Inappropriate loss function: In deep learning models, the loss function is used to perform gradient descent to optimize the model parameters. Choosing an appropriate loss function is particularly important. In existing algorithms, IoU (Intersection over Union) is an important part of the loss function, which determines the accuracy of target localization in detection. However, the sensitivity of IoU for small targets is different from that of medium and large targets. As shown in Figure 1, when the predicted bounding box for small targets and large targets deviate from the ground truth diagonal by 1 and 4 pixels, respectively, the IoU of small targets drops sharply from 0.53 to 0.06, while that of large targets drops from 0.90 to 0.65, with a slower rate of change compared to small targets. To comprehensively and in detail summarize the current status of small target detection methods, this section is divided into three parts: general small target detection methods, small target detection methods in specific field, and commonly used industrial target detection methods, which are also the baseline method of this paper, i.e. YOLO.

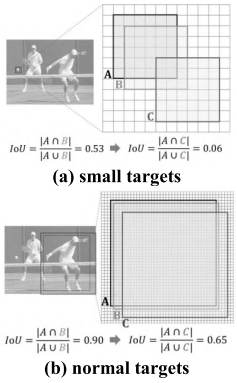


FIGURE 1. The sensitivity analysis about IoU.

A. GENERAL SMALL TARGET DETECTION METHODS

n target detection methods, it is common to start from the aspects of multi-scale features, contextual information, loss functions, etc. Bell et al. [9] proposed the Inside-Outside Net (ION), a target detection network that utilizes information inside and outside the Region of Interest (ROI), integrates context information outside the ROI using spatial recursive neural networks, and extracts feature information using skip pooling. Girshick et al. [10] proposed variable convolution, which improves on the limitations of fixed convolution in extracting spatial information. Li et al. [11] proposed Focal Loss, which dynamically adjusts the contribution of detection results to the loss function based on their confidence, solving the problem of imbalance between positive and negative samples encountered during training of single-stage detectors. Yao et al. [12] proposed SNIPER, a method that solves the problem of long training time and high resource consumption associated with image pyramids in multi-scale training. By appropriately processing context areas around the annotated values at a suitable scale, training speed is greatly improved. Liu et al. [13] proposed DetNet, a backbone network specifically designed for target detection. Compared with ResNet-50, DetNet-59 significantly improves the detection of small targets on the COCO dataset, with AP_50 increasing by 6.4 to reach 66.4. Vu et al. [14] pointed out that most current target detection algorithms use an IoU threshold of 0.5 to determine positive and negative samples. However, using such a wide threshold can lead to a lot of interference, and increasing the threshold can lead to a decrease in detection performance. Therefore, they proposed Cascade R-CNN, a multi-stage target detection architecture that trains using different stages and IoU thresholds in a cascaded manner, avoiding the overfitting problem during training and the mismatching problem during inference.

As the factors limiting the detection performance of small target are increasingly cognized, various methods have been proposed in recent years to improve the accuracy. The following is a comprehensive introduction according to different principles.

1) MULTI-SCALE FEATURE FUSION METHODS

In the target detection task, as the network infers, the features and locational information of the small target gradually get lost in the feature map. The feature pyramid can produce multi-scale features, in which all layers, including the high-resolution layer, have strong semantic information. However, because the multi-scale features in the feature pyramid network are independently computed, the speed is slow. In addition, as deep convolutional networks compute feature hierarchy layer by layer, significant semantic differences are introduced due to the difference in depth. Overall, as the network deepens, it becomes increasingly difficult to preserve the features of small targets, which greatly affects their detection performance. The features of shallow networks have more detailed locational and small target information. Therefore, multi-scale feature fusion of shallow and deep features is an effective solution.

Ma et al. [15] proposed the Feature Pyramid Networks (FPN) for target detection. The FPN structure, as shown in Figure 2, consists of three main parts: the bottom-up path, the top-down path, and the lateral connection. The bottom-up path is the forward propagation process in neural networks. After the continuous convolution operations, the feature maps usually become smaller, achieving the goal of producing multi-scale features. The top-down path uses upsampling operations to extract strong semantic features from high-level feature maps and then fuses them with the original feature maps through the lateral connection. FPN, which combines high-resolution and high-semantic information, was applied to Faster RCNN for small target detection, achieving an average precision of 17.8.

Before feature fusion, FPN performs the 1×1 convolution on the features of different layers to reduce the feature channels. However, since the large semantic gap between features

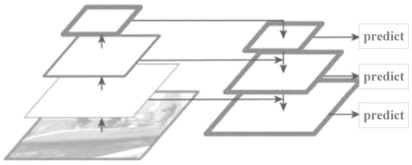


FIGURE 2. FPN.

is not considered, the directly fused features would reduce the ability of multi-scale representation. Specifically, the feature fusion of FPN is performed top-down, leading to the loss of feature information in the highest layer due to channel reduction. After feature fusion, the features of each candidate region are selected from one layer of feature maps based on the scale of the proposal, ignoring other layers that also contain rich information, thus affecting the final detection performance.

PANet [16] is a structure that adds the bottom-up path, the adaptive feature pooling, and the final detection and segmentation block to the backbone of FPN, as shown in Figure 3. For models that use the regression method for prediction, such as R-CNN, FPN, YOLOv3, and YOLOv4, the detailed information in the low-level feature map is important for coordinate regression. However, most models perform coordinate regression on high-level feature maps and lose a lot of detailed information after passing through the backbone network. In FPN, different feature levels are assigned to different sizes of the proposal regions, with smaller proposal regions assigned to lower-level features and larger proposal regions assigned to higher-level features. Although the prediction of FPN is based on multi-level features, each ROI still extracts features based on a single layer.

To address this problem, PANet adds the bottom-up enhancement branch, as shown by the green dashed line in Figure 3, which provides the detailed information required for the coordinate regression, while the original path indicated by the red dashed line is used to transmit semantic information about categories, fully utilizing both low-level and high-level features. In addition, the adaptive feature pooling replaces single-layer features with multi-layer features, and the ROI features obtained from different layers are fused together to obtain the final feature, which is used for subsequent prediction. Specifically, PANet is selected as one of the baseline modules in this paper.

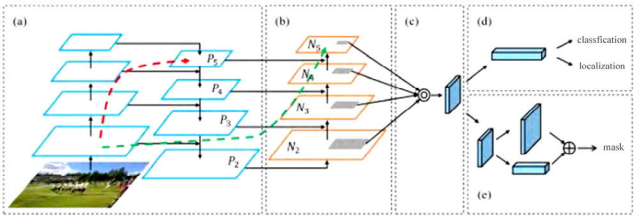


FIGURE 3. PANet. (a) FPN. (b) bottom-up path. (c) adaptive feature pooling. (d) detection branch. (e) fusion layer.

2) DATA AUGMENTATION METHODS

In [17], the data augmentation method is used to improve the accuracy of small target detection. The authors analyzed the detection performance of Mask R-CNN [18] on the MS-COCO dataset and identified two reasons for the poor performance of the model on small targets: 1) there are few images containing small targets and 2) even if the images contain small targets, the proportion is too small. Therefore, the authors over-sampled small target samples and enhanced each image by repeatedly copying and pasting small targets.

During the training phase, images containing small targets were over-sampled to solve the problem of a small number of images containing small targets in the dataset. The sample size was balanced by controlling the number of times the image was copied, that is, the oversampling rate. Since the MS-COCO dataset provides instance segmentation masks, it is convenient to copy from the original location of the target. Therefore, based on oversampling, the copy-and-paste idea was adopted to paste small targets to any other location in the image while generating new labels, and the pasted small targets could be randomly transformed by scaling, rotating, and so on.

The data augmentation method starts with the data level to solve the problem of uneven sample distribution in the dataset. By augmenting the data, the number of small targets in the image is increased, thereby increasing the number of matching anchors and improving the contribution of the loss function calculation during the training phase, resulting in better small target detection. The experimental results show that the accuracy of small target detection was improved by 7.1%.

3) SUPER-RESOLUTION METHODS

In order to improve the localization ability of small-sized images, Jing et al. [19] proposed a new super-resolution method, i.e., Feature Super-Resolution (FSR), which is different from traditional image super-resolution method. Zhang et al. [20] was the first to apply GAN (Generative Adversarial Network) to small target detection tasks, proposing Perceptual GAN. The generator, composed of multiple residual blocks, learns residual representations between targets of different sizes to reduce the gap between small and large targets by enhancing the representation of small targets. The discriminator is composed of the adversarial branch and the perceptual branch. Specifically, the adversarial branch distinguishes between the reconstructed small targets and the large targets, while the perceptual branch is used for classification and regression for target detection. The perceptual branch is first trained with large target features, followed by training the generator with small targets and training the adversarial branch with both large and small targets. However, this method only considers images containing either small or large targets, and adversarial training is difficult for the discriminator to distinguish between the features of large

targets and the small target super-resolution representations output by the generator.

Sun et al. [21] pointed out that small targets are difficult to distinguish from the background or other similar targets due to the lack of feature information, and proposed Multi-Task Generative Adversarial Network (MTGAN). The generator is the super-resolution network that reconstructs small and blurry images by the upsampling operation to restore detailed information in the image. The multi-task module of discriminator includes judging the authenticity of the image, classification, and regression. During training, the loss from classification and regression is backpropagated to the generator, enabling it to reconstruct more details. This method first uses the baseline detector to obtain the target and background of the image. Since the generator performs super-resolution operation on the image, the reconstructed image is not the feature map, so it is necessary to extract features again, resulting in a expensive computational cost.

Deng et al. [22] pointed out the issues of Perceptual GAN network lacking direct supervised signals and MTGAN having excessive computational complexity. They considered that using appropriate high-resolution target features as supervised signals for training the SR (Super-Resolution) model, and the receptive field that matches the input low-resolution features and target high-resolution features can improve the performance of feature super-resolution. The authors added four additional parts on base of the Faster R-CNN base detector: SR feature generator, SR feature discriminator, SR target extractor, and small predictor. As the SR feature generator based on GAN model, it generates high-resolution features with the features extracted by the SR target extractor as the target, under the guidance of SR feature discriminator. The small predictor is used to predict the category and location confidence of small targets, while the original large predictor is used to detect large targets. The authors elaborated on the mismatching problem between the receptive fields of high and low-resolution features and used dilated convolutions to match the receptive field, but did not experimentally explain the matching process, so there may still be the mismatching in receptive fields.

Rabbi et al. [23] proposed an improvement to the S^2A-NET [24] called the S2ANET-SR model, where both the original image and the reduced image are simultaneously inputted to the detection network. To enhance the feature extraction ability of small targets, a SR enhancement module for the reduced image is designed, and perceptual loss & texture matching loss are proposed as the supervision. The mean Average Precision (mAP) on the DOTA dataset reached 74.47%. Yi et al. [25], [26] combined the CycleGAN and Residual Feature Aggregation (RFA) to improve the current SR framework for enhancing detection performance.

The method of small object detection based on super-resolution adds an SR module to the base detector, resulting in the increasing computational cost. The use of the lookup table can reduce the computational cost, but the single-layer based lookup table method limits the scalability and generalization ability of model. Therefore, Ma et al. [27] proposed a serial-parallel lookup table framework to address this problem and achieve efficient image super-resolution framework. In response to the non-local operation in the image SR algorithm, which tends to be global in the receptive field of deep networks, resulting in inaccurate correlation calculations between deep features, and the problem of large computational complexity based on full-image calculation of feature similarity, Li et al. [28] proposed Non-Local Sparse Attention (NLSA), which significantly reduces the computational cost and increases the effectiveness of the non-locality operation. Yang et al. [29] proposed an efficient non-local contrast attention module to address the influence of noise on image super-resolution.

The relative independence of reconstruction and detection algorithms and the computational cost limit their integration to some extent. In addition, the SR network is difficult to train and relies heavily on massive datasets. Therefore, small target detection based on SR still has significant development potential in the future.

B. SMALL TARGET DETECTION METHODS IN SPECIFIC FIELD

Detection of small targets in pedestrian, face, and remote sensing images is an application about specific field. Similar to the general task of small target detection, it also faces challenges such as small scale and limited features, but there are also some differences. Specifically, the distribution of targets in the specific field is usually more dense. In addition, the detection targets for pedestrians and faces are relatively singular, only needing to judge whether the target is the object to be detected, without classification loss. However, detection of small targets in remote sensing images is more complex, as the images are taken from the air angle, and there are difficulties such as target rotation angles.

1) SMALL TARGET DETECTION IN PEDESTRIAN AND FACE

Chen et al. [30] proposed the FSAF (Feature Selective Anchor-Free) module by adding an anchor-free branch after each layer of FPN. Each added branch predicts the same target, and during the backpropagation phase, the layer with the smallest loss is selected to establish the supervision signal, avoiding the defect of anchor-based detectors only dividing the belonging layer based on the target scale during the prediction process. Liu et al. [31] abandoned the anchor and sliding window-based methods and used extracted high-level semantic features to predict the center and scale of pedestrians. Since the scale of small targets is too small, predicting the center point is conducive to locating small targets and is a valuable idea for small target detection. Spyrou et al. [32] proposed a scale matching method for detecting small pedestrians to address the problem of scale mismatch between the data used for detector learning and the data used for network pre-training.

Zhu et al. [33] proposed a scale-balanced face detection architecture to better handle the issue of varying face scales. The used VGG16 occupies about 80% of the inference time, and using a more efficient network can improve detection speed. Christel et al. [34] discussed the problem of detecting small faces from three aspects: scale, resolution, and context. Zhu et al. [35] pointed out that the low IoU between anchors and faces resulting in poor detection performance, so they proposed the EMO (Expected Max Overlapping) score to evaluate the degree of matching between the two and proposed a new anchor design strategy to achieve a high IoU.

Liu et al. [36] used the generator of GAN to reconstruct and deblur high-resolution faces, and the discriminator of GAN was used to identify faces. Smeaton et al. [37] proposed a strategy to dynamically adjust the training weight based on the difficulty of detection. A score representing the difficulty level of each image was assigned during the training phase, and images with high scores were included in a subset for the next round of training.

Unlike small target detection method in remote sensing images, pedestrian and face detection do not need to consider the rotation direction of the target, which reduces the difficulty of detection to some extent. When the general target detector is applied to small and dense face detection tasks, the size of the anchor will not match the receptive field, and small anchors will produce a large number of negative samples during the matching process. Therefore, most scholars tend to study the matching and setting strategies of anchors, and face detection has achieved good results now.

2) SMALL TARGET DETECTION IN THE REMOTE SENSING IMAGE

Remote sensing images are captured from a high altitude perspective and have complex spatial scenes with various target types. Target detection in the remote sensing images faces difficulties such as small and dense target scales, complex backgrounds, and arbitrary distribution directions. Ding et al. [38] proposed a Dataset of Object deTection in Aerial images, i.e., DOTA. Pang et al. [39] proposed a unified self-enhanced network called the Remote Sensing-based Convolutional Neural Network (R2-CNN), which consists of the lightweight network Tinny-Net, the global attention module, the classifier, and the detector. The Tinny-Net makes the network highly efficient in terms of computation and memory consumption, and the global attention module provides strong robustness against false positives. Yang et al. [40] proposed a feature fusion structure to solve the problem of small targets by exploring anchor sampling angles and feature fusion.

Li et al. [41] proposed a new semantic representation method to improve the performance of detecting remote sensing images. They first designed an enhanced feature pyramid network to better extract visual features with hierarchical differences, then introduced semantic segmentation to guide the detection of horizontal proposals, and finally proposed an ROI module that fuses multiple-layer features to learn target-based semantic representations on the existing features.

Yang et al. [42] observed that targets in aerial images are clustered, so they integrated clustering with target detection and proposed the ClusDet network. Han et al. [43] incorporated the rotation-equivariant network into the detector, enabling it to predict the direction of the target when extracting rotation-invariant features. The proposed rotation-equivariant detector ReDet can solve the problem of arbitrary distribution directions of aerial targets.

Qin et al. [44] proposed a multi-head rotated target detector called MRDet to predict the classification confidence, location, scale, and direction of the final bounding box separately. They divided the detection task into multiple sub-tasks, and each detection head was specially designed to learn the features that were most suitable for the corresponding task.

Yi et al. [45] extended the horizontal landmark-based target detector to the directional target detection task to address the severe imbalance problem between positive and negative anchors encountered by current anchor-based two-stage detectors when detecting targets in aerial images with arbitrary and densely arranged directions. The authors first detected the center landmark of the target and then regressed the bounding box aware vectors (BBAVectors) to capture the directional bounding box.

Wei et al. [46] applied Transformers to small target detection and proposed CG-Net (Calibrated-Guidance) to enhance the relationship between channels in a feature transformer manner. This method can adaptively determine the calibration weights of each channel, and by aggregating all weighted channels together, it can represent each channel again. Wang et al. [47] proposed a visual model for remote sensing tasks based on ViT (Vision Transformer) and a new rotation-variable window attention to replace the full attention in the original Transformers. This method learns better target representations by extracting rich context from the generated different windows.

Shamsolmoali et al. [48] introduced the image pyramid into SSD (Single Shot MultiBox Detector) and proposed IPSSD (Image Pyramid Single-Shot Detector) for detecting small targets in remote sensing images. Although image pyramids can extract more semantic features, they inevitably bring additional computational and memory costs.

Target detection in the remote sensing images belongs to the scope of specific small target detection. The methods in the above literature include commonly used techniques such as attention mechanism and feature fusion, as well as unconventional methods such as introducing Transformers and using multiple detection heads to predict classification confidence and location separately. Overall, remote sensing

image detection is a rapidly developing field with vast potential for further growth.

C. YOLO FAMILY

Target detection based on deep learning can be divided into two categories: two-stage detection and single-stage detection [49]. The former is a coarse-to-fine process, while the latter is an end-to-end one-step process [50]. Generally, the localization and classfication accuracy of two-stage detection is higher, while the speed of single-stage detection is faster. Typically, single-stage detection attempts to directly classify each RoI as either background or target [51]. That is, it can directly give the category probability and location coordinates of the target through the single stage, and the typical representatives include YOLO family [52].

1) FUNDAMENTAL THEORY

The basic framework of YOLOv1, as shown in Figure 4, first adjusts the size of the input image to 448 × 448, and then sends it to the backbone structure to extract features. Then, the network predicts the results and achieves end-to-end target detection. YOLOv1 abandons the traditional sliding window technique. It divides the input image into $S \times S$ grids, and each grid is responsible for detecting the targets whose centers fall within that grid. Each grid predicts B bounding boxes and their confidence scores. The confidence score includes the probability of the bounding box containing an target and the accuracy of the bounding box. Each bounding box predicts 5 elements, i.e., (x, y, w, h, c), representing the location, size, and confidence score of the bounding box. Each grid predicts $(B \times 5 + C)$ values, where C is the number of categories. Then, the network prediction is performed using the Non-Maximum Suppression (NMS) algorithm. Subsequent models in YOLO family have inherited this basic idea.

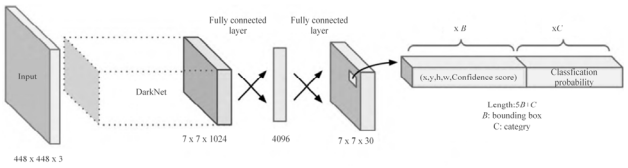


FIGURE 4. YOLOv1 architecture.

2) EVOLUTION OF BACKBONE NETWORK

Detectors typically consist of two parts: the Backbone network, which is the basic network used for extracting features and is usually pre-trained on the ImageNet dataset, and the Head for predicting target categories and bounding boxes [53], [54]. In recent years, the Neck has been constructed between the Backbone and Head to aggregate different feature maps. The following will provide a detailed analysis of the evolution of the backbone network in YOLO family.

YOLO V1. YOLOv1 [52] uses the Backbone network similar to GoogleNet [55], consisting of 24 convolutional layers and 2 fully connected layers. It is pre-trained on ImageNet dataset and then transferred to the detection task, and validated on the VOC (Visual Object Classes) dataset [56].

YOLOv1 divides the input image into the grids of 7×7, and predicts two bounding boxes for each grid, resulting in a total of $7 \times 7 \times 2$ bounding boxes. It can detect up to 49 targets, which makes it less effective at detecting dense and small targets.

YOLO V2. YOLOv2 [57] uses the VGG network as the reference and constructs a new Backbone network called Darknet-19 based on YOLOv1. YOLOv1 directly predicts bounding boxes using fully connected layers, which causes inaccurate localization due to significant loss of spatial information. Therefore, YOLOv2 introduces anchors to replace the fully connected layers in v1 for predicting bounding boxes. Meanwhile, YOLOv2 resizes the input to 416 × 416 and obtained the feature map of 13 × 13 with odd dimensions, resulting in only one center for each grid. This center point is used to predict the target falling into that location, making it easier to detect that particular class of targets. Figure 5 shows the method in YOLOv2 used to predict bounding boxes.

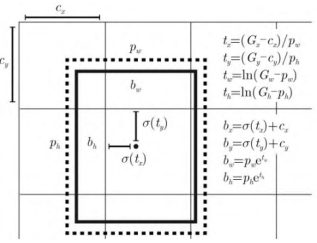


FIGURE 5. Bounding box with scale prior and location prediction. The dashed rectangle represents the anchor, while the solid rectangle represents the predicted bounding box obtained by offsetting the anchor through the network. In addition, (cx, cy) represents the coordinates of the upper-left corner of the grid, (pw, ph) represents the width and height of the anchor, and (tx, ty) and (tw, th) represent the center offset and width-to-height ratio of the predicted bounding box, respectively. The ground truth coordinates in the feature map are denoted as (Gx, Gy, Gw, Gh), and (bx, by, bw, bh) represents the final predicted bounding box about target. The conversion process from the proposal bounding box to the predicted bounding box is shown in the right-hand side, where σ is the sigmoid function used to scale the predicted offsets to the range of 0 to 1, accelerating the convergence of the network.

YOLOv2 [57] proposes a groundbreaking method that jointly training classification and detection, extending detection to targets with a lack of samples. This work significantly improves prediction accuracy while maintaining the advantage of fast inference.

YOLO V3. The basic network of YOLOv3 [53] is Darknet-53, which borrows the residual structure of ResNet [58] to deepen the network structure while preventing the problem of convergence difficulty caused by the gradient explosion. During forward propagation process, the pooling layer and fully connected layer are removed, and the size of the tensor is changed by changing the stride of the convolution kernel. Similar to v2, Darknet-53 reduces the output features to $1/2^8$ of the input, so it is usually required that the resolution of input image be a multiple of 32. At the same time,

YOLOv3 uses tensor concatenation to expand the dimension of the tensor and extract more information. Specifically, the intermediate layer and the later layer of Darknet-53 are concatenated after upsampling. Darknet-53 has 53 convolutional layers from the 0th to the 74th layer, and the rest are residual layers [53]. The 75th to 105th are the feature fusion layers of YOLOv3, which adds multiscale detection (equivalent to the Neck) using 3 scales. The output of each scale is 52×52, 26×26, and 13×13, respectively, for detecting small, medium, and large targets, in which each scale predicts 3 anchors.

In summary, the number of predicted anchors in YOLOv3 is more than 10 times that of YOLOv2, and they are performed at different scales, so the overall precision of detecting small targets have been greatly improved. Therefore, it has become one of the milestone architectures in single-stage detection.

YOLO V4. YOLOv4 [54] summarizes various improvement methods after v3, which are divided into free and special packages. The former represents modules that improve training without affecting inference speed, while the latter represents modules that have little impact on inference time but offer higher performance return, such as the CSP (Cross Stage Partial) [59] structure used in the Backbone, which maintains high inference speed while still having high accuracy. Meanwhile, YOLOv4 is more suitable for training on a single GPU.

Bochkovskiy et al. [54] found that when the model is optimal for classification, it is not necessarily optimal for detection. For example, the classification accuracy of CSPResNeXt-50 is higher than that of CSPDarknet-53, but the latter has higher detection accuracy. Therefore, YOLOv4 chooses CSPDarknet-53 as the backbone network.

Regarding the Backbone, the overall architecture of YOLOv4 is the same as YOLOv3, but improvements have been made to each substructure. Figure 6 shows two network structures: Darknet-53 and CSPDarknet-53 [59]. The black color represents Darknet-53, and the CSPDarknet-53 network only needs to be replaced with the structure in the red box, and the filter values are changed to the red values in parentheses. YOLOv4 removes the last pooling layer, fully connected layer, and Softmax layer, and its Backbone has five CSP modules [54].

For the Neck, YOLOv4 introduces the Spatial Pyramid Pooling (SPP) and PANet modules. SPP significantly increases the receptive field and separates important contextual features without reducing running speed. PANet replaces FPN used in YOLOv3 for parameter aggregation and uses tensor connections instead of the original short connections.

For the Head, YOLOv4 inherits the multi-scale structure from YOLOv3 for prediction.

YOLO V5. YOLOv5 [60] has a similar basic structure to YOLOv4, with the main difference being the scaling of different channel sizes. Based on model size, YOLOv5 offers five different models: YOLOv5-n/s/m/l/x.

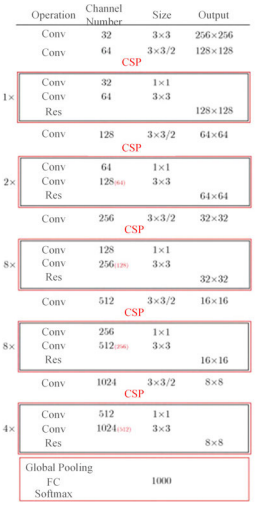


FIGURE 6. Detailed information of Darknet-53 and CSPDarknet-53.

As mentioned above, it can be seen that the methods in the YOLO family directly divide the image into several regions and predict the bounding box and probability for each region at the same time, which greatly improves the detection speed.

III. THE PROPOSED METHOD

A. OVERALL ARCHITECTURE

Compared with the background region, small targets are very small in size and lack self-information. Directly inputting images containing small targets into YOLOv5 will cause the network to treat them as common targets and ignore their special characteristics. Firstly, inspired by [7] and [61], this paper improves the channel-spatial attention mechanism under a single receptive field and proposes a multi-branching blended attention block, i.e., MBAB, combined with the multi-branch-ConvNet structure. Compared with [7], the proposed module can more effectively mine the feature information of small targets with a tiny increase in computational cost and dynamically allocate blended attention weights according to the contribution of feature maps of different scales to small targets. Then, an improved feature extraction module CMBAB is introduced at the end of the Backbone to enhance the feature extraction capability of the core network, and MBAB and CMBAB are introduced in the up-sampling and down-sampling operations of the Neck's multi-scale fusion to enhance the feature expression ability about small targets. Finally, the P2 detection branch is added to the PANet for detecting small targets. In summary, the overall architecture of the MBAB-YOLO is shown in Figure 7.

B. MULTI-BRANCHING BLENDED ATTENTION BLOCK-MBAM

Most deep learning-based target detection methods use ConvNets, but different convolution kernels have different sensitivities to targets of different sizes. Specifically,

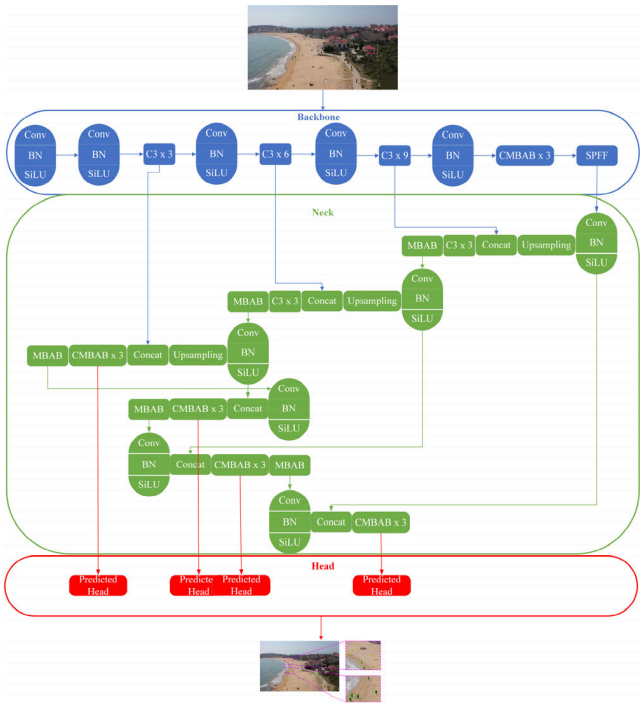


FIGURE 7. The proposed YOLO architecture.

Szeqedy et al. [62] proposed GoogleNet, which achieved certain superiority by using the Inception structure, consisting of four network blocks with different convolution kernels. Later, Xie et al. [63] proposed ResNeXt, which introduced group convolution in the bottleneck of ResNet and used a multi-branching structure in the base architecture, demonstrating the effectiveness through extensive experiments. SE-Net enhanced effective target features and suppressed background information by adding channel-wise attention mechanisms to adaptively re-calibrate features. SK-Net (Selective Kernel Network) [64] used two different convolutional kernel branches, also introduced channel-wise attention mechanisms for fusion features, and then adaptively split the branch network for re-calibration. ResNeSt [61] improved SK-Net by using different n convolution kernels and using the dilated convolution to share computations. After introducing the channel-wise attention block, the features were re-calibrated with n attention mechanisms for different receptive fields. CBAM (Convolutional Block Attention Module) redefined channel-wise attention block by adding the sum of the mean and maximum values between channels, and introduced spatial-attention in the same way. The concatenated and blended channel and spatial attention mechanisms were used to re-calibrate the feature map. Experimental results showed that its effect was superior to that of a single attention mechanism. Inspired by these methods, this paper improves the blended attention block and combines the multi-branch-ConvNet structure with the blended attention mechanism, and verifies the effectiveness of the proposed method through extensive experiments.

The main design inspirations of the proposed MBAB in this paper is as follows: First, the input feature map F is processed with different convolution kernels to obtain multiple branches, and then the multiple branches are summed to obtain the blended feature map $F(b)$. Then, the attention weight matrices X_{SA} and X_{CA} are separately calculated along the spatial and channel dimensions of the blended feature map $F(b)$, and the weights are fused along the spatial and channel dimensions to obtain the mixed attention weight matrix X_{BAB}. The blended attention weight is then redistributed according to the contribution of each branch, and the feature map under different convolution kernels is re-calibrated. Finally, the weighted feature maps are summed to obtain the blended attention-weighted feature map. For small targets, during the multi-branching stage, with the training of the network, the features of small targets will be assigned different weights on branches with different receptive fields, where positive targets will be assigned larger weights while negative targets will be assigned smaller weights. Through this multi-branching structure, the model will focus more attention on learning effective features, thereby enhancing its generalization ability. Inspired by [7], two convolutional attention modules, channel-wise attention (CA) block and space-attention (SA) block, are designed.

The structure of CA is shown in Figure 8, and its calculation is shown in (1):

$$X_{CA} = Sig \bmod \{f_c\ (AdaAvgPooling(F)) + f_c(AdaMaxPooling(F))\} \quad (1)$$

in which, the size of the input feature map F is $N\times C\times H\times W$; Sigmod is the activation function; *AdaAvgPooling* is the global adaptive average pooling; *AdaMaxPooling* is the global adaptive max pooling; and f_c is the fully connected network. First, global adaptive average pooling and max pooling are applied to the feature map F, then the two obtained channel weights are passed through the fully connected layer f_c (Conv & Relu & Conv), and finally the two different weights of the fully connected are summed and Sigmod is activated to obtain the channel attention X_{CA}, whose size is $N\times C\times 1\times 1$.

The structure of SA is shown in Figure 9, and its calculation is shown in (2):

$$X_{SA} = Sig \bmod \{f_c\ (cat(mean(F), \max(F)\ ))\} \quad (2)$$

in which, the input feature map F has a size of $N\times C\times H\times W$, where *mean* is the mean function, *max* is the max function, and *cat* is the matrix concatenation function. Firstly, the mean and maximum are computed spatially for the feature map F, resulting in the size of $N\times 1\times H\times W$. Then, the weight matrices for the mean and maximum are concatenated spatially to obtain the double-channel spatial attention weight. Finally, double-channel composite spatial attention X_{SA} is obtained through the 1×1 fully connected convolution layer, resulting in the size of $N\times 1\times H\times W$.

In reference to [7], after the channel-wise attention is applied, the feature space is re-calibrated, then

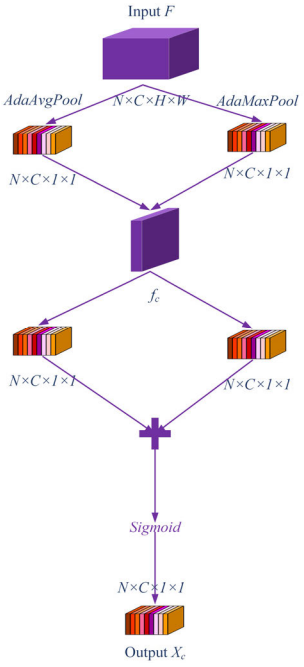


FIGURE 8. Channel-wise attention block.

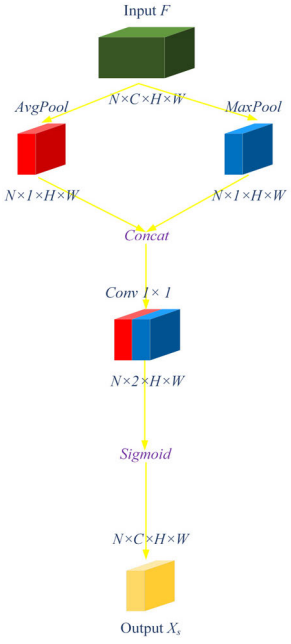


FIGURE 9. Space-attention block.

space-attention is concatenated and the feature space is re-calibrated again, resulting in two rounds of attention-weighted feature maps. However, the output is the weighted feature map that is not conducive to combining the multi-branch-ConvNet. In addition, the concatenation operation is simple and effective, but it does not adequately consider the impact of reasonable connection methods on small targets. To improve the CBAM network structure and facilitate its

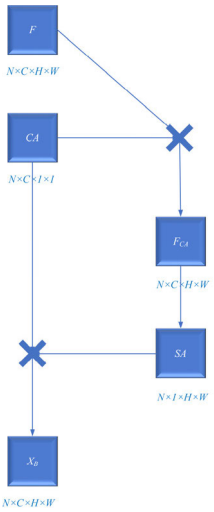


FIGURE 10. Proposed blended attention block.

combination with the multi-branch-ConvNet without increasing computational complexity, we propose a blended attention block (BAB), as shown in Figure 10.

The calculation of BAB is shown in (3):

$$X_{BAB} = mul(CA(F), SA(F \otimes CA(F))) \tag{3}$$

in which, the input feature map F has a size of $N\times C\times H\times W$, CA is channel-wise attention block, output size is $N\times C\times 1\times 1$; SA is space-attention block, output size is $N\times 1\times H\times W$; *mul* is matrix multiplication, and X_{BAB} is the uncalibrated blended attention weights of CA and SA, with the output size of $N\times C\times H\times W$.

The multi-branching attention module in [61] only uses channel-wise attention and has certain effect on feature extraction for small targets, but lacks consideration of spatial dimensions and is not comprehensive. Therefore, the blended attention module is introduced into the multi-branching network in this paper, and the multi-branching blended attention block MBAB is proposed, whose network structure is shown in Figure 11.

Assuming that the input feature map F has a size of $N\times C\times H\times W$ and the number of split branches is S, a series of mappings $\{F_1, F_2, \ldots, F_S\}$ are obtained through transformations with different convolution kernels. In addition, for each mapping, the element-wise sum fusion is performed to obtain the multi-branching blended feature map, denoted as F_{MB}, with the formula shown in (4):

$$F_{MB} = \sum_{i=1}^{S} F_i \tag{4}$$

Then, the F_{MB} is used as the input of the proposed BAB, and the output is the multi-branching blended attention weight matrix, denoted as X_{MB}, where $X_{MB}= BMB(F_{MB})$, with a size of $N\times C\times H\times W$. X_{MB} is then divided into d

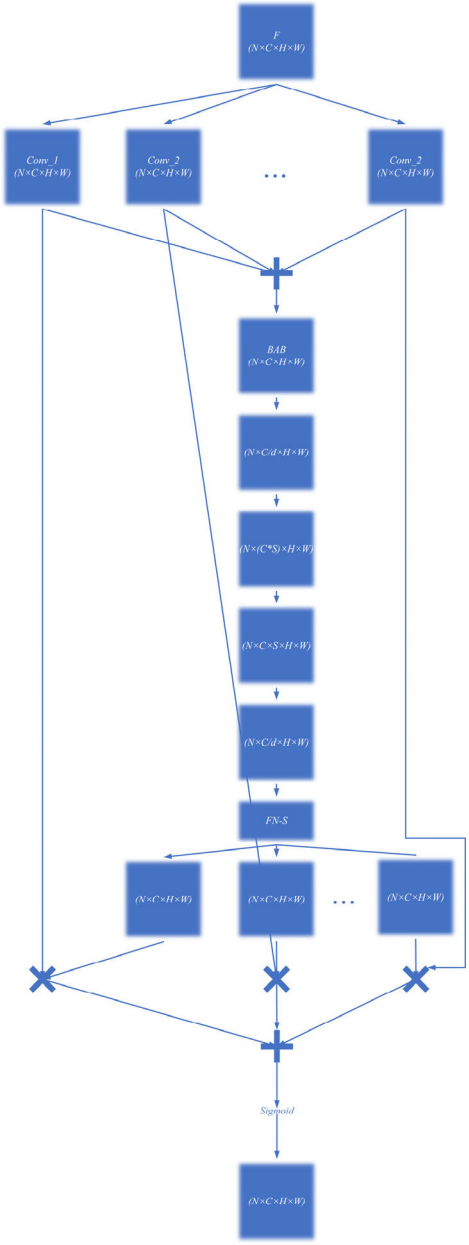


FIGURE 11. Proposed MBAB.

groups (d is the hyperparameter), and each group is called a base, denoted as X_{MBd}, where $X_{MBd} = f(X_{MB})$, and f is the 1×1 convolution with a size of $N \times C/d \times H \times W$. Therefore, the multi-branching blended attention weight matrix now has $S \times d$ groups. Then, the global context attention weight matrix is fused to obtain S groups of blended attention weight matrix, denoted as X_{MBS}, where $X_{MBS} = f(X_{MBd})$, and f is the 1×1 convolution with a size of $N \times C * S \times H \times W$. Finally, the blended attention weight is allocated to each branch. In [61], the allocation weight is calculated using softmax function. On the contrast, in order to reduce computation complexity, the Faster Normalization (FN) method is used instead of the original softmax method, with negligible additional cost. The FN is shown in (5):

$$O_i = \frac{W(F_i)}{\gamma + \sum_{i=1}^{S} W(F_i)}, \quad i \in \{1, 2, 3 \ldots S\} \tag{5}$$

in which, O_i represents the contribution of different receptive field branches to the overall blended attention size. The overall blended attention size reassigns the blended attention of each branch based on O_i, where γ is the constant, usually taken as a very small value of 0.00001 to prevent regularization failure caused by a denominator of zero. After the weights of each branch are allocated, the multi-branching feature map is re-scaled and the corresponding elements are summed to obtain the feature map F_{out} weighted by multi-branching blended attention. F_{out} is shown in (6).

$$F_{out} = \sum_{i=1}^{S} (FN(X_{MBS}^{i}) \otimes F_i) \tag{6}$$

in which, S is the number of branches; i is the rescaled branch; and F_i is the i-th branch. The final output size is $N \times C \times H \times W$.

C. FEATURE EXTRACTION MODULE-CMBAB

The core module C3 for feature extraction in YOLOv5 mainly uses the CSPNet network structure to stack Bottleneck residual blocks. To improve its structure, the 3×3 convolution in the residual block is replaced with the multi-branching blended attention block, i.e., MBAB. To reduce the number of parameters, the modified residual blocks are no longer stacked, and the feature extraction module called CMBAB, based on CSPNet and MBAB, is proposed. Its network structure is shown in Figure 12.

Firstly, assuming the input feature map F_{in} has a size of $N \times C \times H \times W$. Then, after 1×1 convolution, it is used as the input of the residual block and denoted as F_{res-in}. After passing through the residual connection with MBAB, the output feature map is denoted as $F_{res-out} = add(F_{res-in}, MBAB(f(F_{res-in})))$, where add represents element-wise summation, and f is the 1×1 convolution. Finally, by combining with the CSPNet structure, CMBAB is obtained with an output denoted as $F_{out} = f(cat(f(F_{in}), F_{res-out}))$, where cat denotes matrix concatenation operation. In terms of module size, when the C3 stacks more than one Bottleneck residual block, the number of parameters in CMBAB is lower than that of C3. In feature extraction and processing for small targets, CMBAB can utilize the proposed MBAB mechanism to obtain more abundant features of small targets than C3, thereby enhancing feature representation and improving detection performance.

IV. EXPERIMENT

A. DATASET

The dataset used in the experiment is the TinyPerson dataset, which is a small target dataset with high-quality annotations [66]. The images in this dataset are mostly taken by

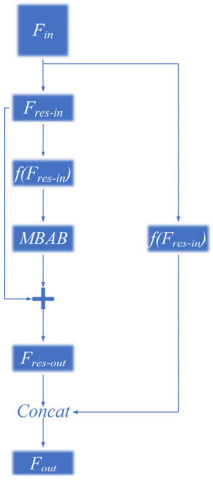


FIGURE 12. Proposed CMBAB.

aerial photography at a long distance with the large background, which is the typical small target detection scenario. The TinyPerson dataset contains two categories, i.e., earth person and sea person, with a total of 1,610 images, in which 794 in the training set and 816 in the testing set, including 72,651 human target annotations.

B. EVALUATION METRICS

In order to evaluate the effectiveness of the model, this paper uses Average Precision (AP), mean Average Precision (mAP), Giga Floating-point Operations Per Second (GFLOP/s), and Frame Per Second (FPS) as the evaluation criteria.

Assuming that the number of correctly detected targets in the results is TP, the number of incorrectly detected targets in the results is FP, and the number of targets that were not detected among the correct targets is FN, the precision rate P (Precision) and recall rate R (Recall) are defined by (7) and (8):

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

To establish a two-dimensional coordinate axis with the x-axis as recall rate and the y-axis as precision rate, simultaneously draw the Precision-Recall (PR) curve, and the area surrounded by the PR curve is the size of the AP, as shown in (9):

$$I_{AP} = \int PdR \tag{9}$$

mAP is the mean average precision of multiple categories, where n represents the total number of categories, and the formula is shown in (10):

$$I_{mAP} = \frac{1}{n}\sum_{i=1}^{n} I_{AP_i} \tag{10}$$

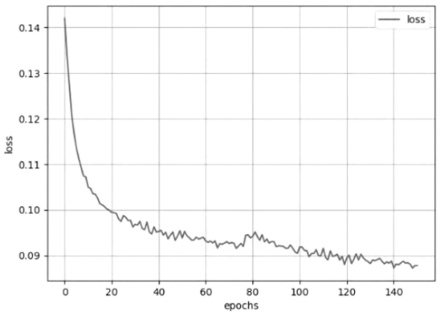


FIGURE 13. Training loss.

GFLOP/s is the floating-point operation per second, usually used to measure the computational complexity of the model.

Frame rate represents the number of images that can be detected per second (unit: frame/s), which is used to measure whether the algorithm has real-time performance. It is generally believed that FPS greater than 30 frame/s indicates real-time detection effect.

AP50 and mAP50 represent the average precision and mean average precision when IoU threshold is 0.5. Generally, the higher the IoU, the larger the intersection between the predicted target and the ground truth, and the closer to the expected target. At this time, the larger the detection accuracy indicates the stronger the prediction ability of model. It is obvious that small targets have fewer pixels, and if a larger IoU is used, the accuracy will be very low, which cannot measure the effect of small target detection algorithms well. Therefore, this paper chooses the compromise solution of IoU of 0.5.

C. EXPERIMENT SETTINGS

The hardware environment for the experiment consists of an Intel Core i7-10750H CPU @ 2.60GHz, 16GB RAM, and NVIDIA GeForce GTX 1660Ti GPU. The software environment consists of Windows 11 system, python3.7, PyTorch 1.8.3, and cuda11.2. Figure 13 shows the curve of regression loss during training. The batch size is set to 4, the training is conducted for 150 epochs, with the first three epochs being warm-up. The optimizer used is SGD with an initial learning rate of 0.01, momentum of 0.937, and learning rate decay using the cosine strategy. As can be seen from Figure 13, the regression loss during training can smoothly decrease and achieve the desired effect. Except for necessary improvements, all hyper-parameters for the models in this experiment are set to default (not necessarily optimal) and training, validation, and testing are performed under this setting.

D. ABLATION EXPERIMENTS

Since the default input size of YOLOv5s is 640×640, but the targets in the TinyPerson dataset are small targets in aerial images with distant backgrounds, and the image sizes are much larger than the default size. Obviously, the larger the

input resolution, the more advantageous it is for detecting small targets, but larger resolutions will result in more expensive computational costs, and the FPS for detecting images will also be lower.

Therefore, this paper tests YOLOv5s and proposed method at different resolutions, and the test results are shown in TABLE 1. Based on the TinyPerson dataset, when the resolution of YOLOv5s trained is 960 × 960 and 1280 × 1280, the mAP_{50} has increased by 9.07% and 15.65% respectively compared to 640 × 640, and the FPS is 122 frames per second at a testing resolution of 1280 × 1280. As a contrast, MBAB-YOLO has increased the mAP_{50} by 7.66% and 13.91% respectively compared to 640 × 640 at resolutions of 960 × 960 and 1280 × 1280, and the FPS is 74 frames per second at a testing resolution of 1280 × 1280. This indicates that increasing the image resolution of the model can improve its accuracy under the same target detection network structure. However, increasing the resolution will increase the computational cost several times, and the training time will also increase significantly. For example, under the experimental configuration of MBAB-YOLO, if 1280 × 1280 resolution is used as the network input, one epoch of training takes 26 minutes, and the computation complexity is 74.4GFLOP/s, while for input at a resolution of 640 × 640, it only takes about 9 minutes for one epoch of training and the computation complexity is 19.9GFLOP/s. Correspondingly, the size of the generated weight file will also increase, making it more difficult to deploy the model. However, if different models trained with different resolutions are used for the same network structure, and the input resolution for detection is the same, it will not affect FPS, which means that the input resolution of the model can be appropriately increased during training to improve detection accuracy. In addition, excessively high resolutions can also cause overfitting, so blindly increasing the resolution is not recommended.

TABLE 1. Influence of different input resolutions on ablation experiments. the value in the input resolution field represents width only, and height equals width. in fps^{1280}, 1280 means that the resolution of the image during the testing is 1280 × 1280.

Methods	YOLOv5s			MBAB-YOLO		
Input_resolutions	640	960	1280	640	960	1280
Num_params(10^6)	7.02	7.02	7.02	7.37	7.37	7.37
Sizes(MB)	56.8	57.0	57.3	60.5	61.4	62.6
	1	2	1	9	4	2
GFLOP/s	15.8	33.9	57.3	19.9	42.6	74.4
mAP_{50}(%)	32.2	41.3	47.9	38.1	45.8	52.0
	7	4	2	6	2	7
FPS^{1280}	122	122	122	74	74	74

To validate the effectiveness of proposed MBAB and CMBAB, and to investigate the impact of additional small target detection head on the results, experiments are conducted to evaluate the effects of different modules on the results. As usual, YOLOv5s is used as the baseline model, with the training resolution of 1280 × 1280 and the testing resolution of 1280 × 1280. A total of 150 epochs are trained with pre-trained weights to accelerate training process, and the experimental results are shown in TABLE 2.

TABLE 2. The results of the ablation experiments.

Methods	α	β	γ	δ
Baseline	T	T	T	T
P2	F	T	T	T
MBAB	F	F	T	T
CMBAB	F	F	F	T
Layer count	270	328	496	587
Num_params(10^6)	7.02	7.17	7.62	7.37
Sizes(MB)	57.31	60.63	64.43	62.62
GFLOP/s	57.27	65.39	76.16	74.4
mAP_{50}(%)	47.92	49.70	51.71	52.07
FPS^{1280}	122	91	78	74

1) IMPACT OF ADDITIONAL DETECTION HEAD

From TABLE 2, it can be seen that α is the baseline model without improvement, and β adds a P2 detection head on base of the baseline model. Since the P2 detection layer has more shallow information, it is more favorable for small target detection. The experimental results show that compared with α, the number of layers in β increases from 270 to 328, GFLOP/s increases from 57.27 to 65.39, and the parameter quantity increases from 7.02M to 7.17M, but the mAP_{50} for small targets increases by 1.78%, and the FPS still meets the requirements of real-time detection. Therefore, it is worthwhile to increase a small amount of computation to achieve better small target detection performance.

2) IMPACT OF BLENDED ATTENTION BLOCK

MBAB can adaptively combine the blended attention mechanism to adjust different receptive fields that are more suitable for small targets, and thereby obtain fully weighted and re-calibrated feature maps for small targets. γ introduces MBAB in front of the SPPF (Spatial Pyramid Pooling-Fast) layer of Backbone and the final part of the up-sampling and down-sampling in the Neck, respectively. Compared with β, the number of layers, GFLOP/s, and parameter quantity of γ increase by 168, 10.77, and 0.45M, respectively, but the mAP_{50} increases by 2.01%. In addition, δ is based on γ, and replaces the original C3 module in CSPNet structure, which is not attention-weighted, with the CMBAB structure, before the downsampling of Neck inputting Head. Compared with γ, δ increases the number of layers by 91, decreases GFLOP/s by 1.76, decreases parameter quantity by 0.25, and at the same time, increases mAP_{50} by 0.36%. The experimental results show that adding the MBAB mechanism and replacing C3 with the lighter CMBAB structure in the feature extraction can enable the model to obtain more abundant small target features.

In summary, compared with the baseline α, proposed δ in this paper has increased the mAP_{50} by 4.15%, significantly improving the detection accuracy of small targets, and the

TABLE 3. Contarst experimental results.

Methods	CBAM	YOLOX-S	PP-YOLO-S	DETR	YOLOv7-tiny	YOLOv5s7	MBAB-YOLO
Num_params(10^6)	7.23	9.01	7.91	41.00	6.02	7.02	7.37
Sizes(MB)	60.41	212.23	59.16	123.65	48.58	60.28	62.62
GFLOP/s	64.56	92.99	63.37	86.01	47.38	60.28	74.42
mAP_{50}(%)	50.61	47.61	48.23	46.16	45.23	50.02	52.07
FPS^{1280}	77.02	69.98	117.08	27.90	131.21	63.29	74.07

FPS reaches 74 frame/s, showing the worthy real-time detection capabilities.

E. CONTRAST EXPERIMENTS

Based on the TinyPerson dataset, contrast experiments are conducted about MBAB-YOLO and several benchmark methods for real-time small target detection, i.e., CBAM, PP-YOLO [67], DETR [68], YOLOv7 [69], YOLOX [70], and YOLOv5, in which YOLOX and PP-YOLO use the s version, YOLOv7 uses the tiny version, and YOLOv5s uses the newer 7.0 version. The experimental results are shown in TABLE 3. According to TABLE 3, among the compared algorithms, YOLOv5 with CBAM has the highest mAP50 of 50.61%, which is 1.46% lower than that of MBAB-YOLO. However, the other parameters in our method are similar to CBAM, ensuring a stable balance between detection accuracy, speed, parameter quantity, and model size. YOLOv7-tiny has the fastest detection speed of 131.21 frames per second, with the smallest parameter quantity, model size, and GFLOP/s, but its mAP50 is 6.84% lower than that of MBAB-YOLO. Obviously, MBAB-YOLO focuses more attention on small targets, dynamically re-calibrating small targets in feature maps of different scales, not only improving detection accuracy but also ensuring speed and real-time performance, which has the worthy advantages in real-time detection of small targets tasks.

F. VISUALIZATIONS

In order to intuitively verify the influence of the proposed multi-branching blended attention block on the features of small targets, as well as the impact on the final detection of small targets, representative images from the TinyPerson testing set were used for validation. Figure 14 shows the baseline YOLOv5 model without BAB, and Figure 15 shows the YOLOv5 model added BAB, with Grad-CAM [71] used for heatmap visualization. From the comparisons of the two figures, we can see that the proposed blended attention block can focus the features of small targets more effectively at different angles when the number of targets is large and the targets are small, which is shown in the Figure 15 as clearer target boundaries and obvious color differences from the environment.

Figure 16 and Figure 17 show the practical detection results of YOLOv5 before and after adding BAB respectively, where the red box represents the human label detected by the model, and the yellow box highlights the differences. From the Figure 16 and Figure 17, it can be seen that under different conditions, such as on land or on sea, during the day or in the evening, the baseline model misses small targets, while the addition of proposed blended attention can better detect small targets. In summary, the proposed multi-branching blended attention block can effectively improve the detection performance of small targets.

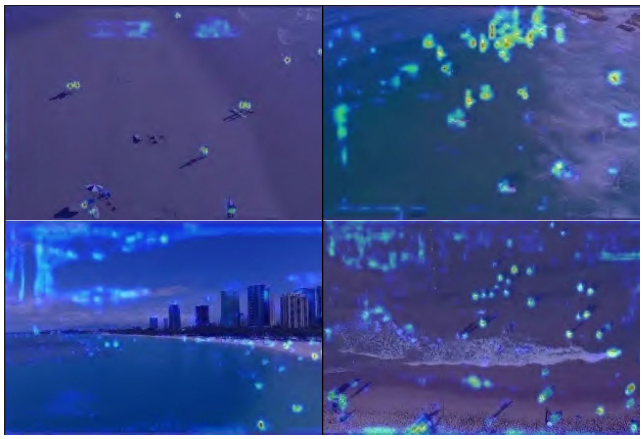


FIGURE 14. Heat map visualization before BAB is added.

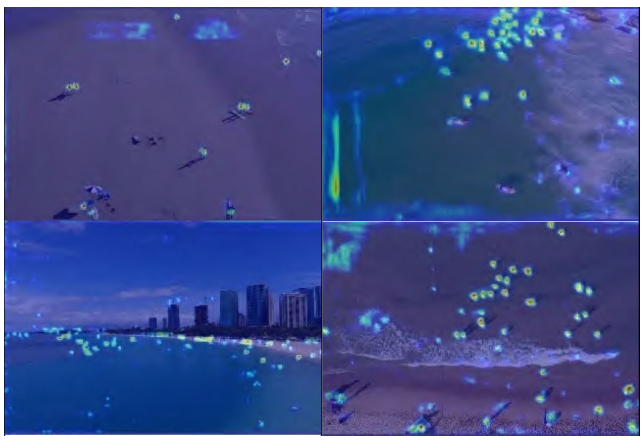


FIGURE 15. Heat map visualization after BAB is added.

V. FUTURE WORKS

The main approach proposed in this paper to improve the YOLOv5s model for small target detection is the multi-branching blended attention block, and it does not

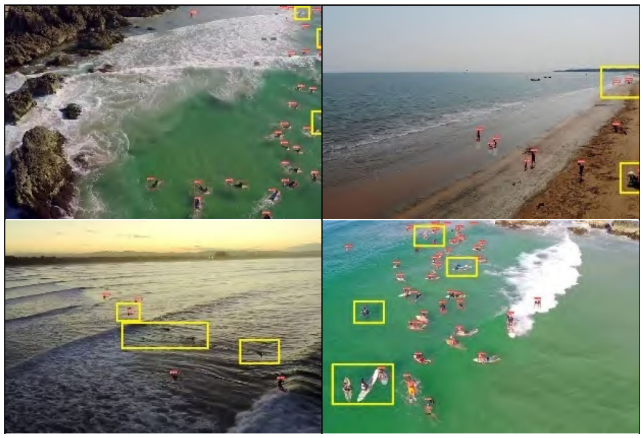


FIGURE 16. Detection visualization before BAB is added.

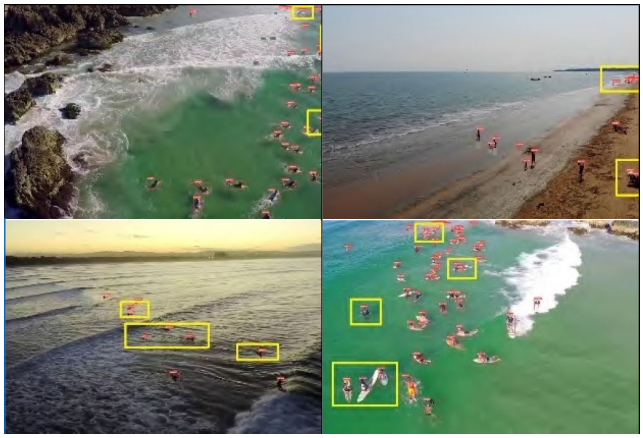


FIGURE 17. Detection visualization after BAB is added.

consider combining with other related methods (such as data augmentation and self-attention mechanism, etc). Therefore, in future research, our study can be conducted on how to combine MBAB-YOLO with more advanced methods to achieve high-performance small target detection. On the whole, in the development of target detection method, people have gradually discovered problems and proposed corresponding solutions. The accuracy of small target detection has gradually improved, but there is still a lot of room for improvement.

A. MULTI-SCALE FEATURE FUSION

After several years of research, the multi-scale feature fusion in ConvNets has achieved good results in the effectiveness and efficiency, but there is still a lack of the mathematical principle and the interpretability. Currently, most feature fusion structures rely on empirical design and experimental improvements. In other words, feature fusion structures, even feature extraction models, can be seen as numerical fitting results of a large amount of data, lacking reasonable explanations. Therefore, there is still a large space for interpretation in the subsequent development. In addition, Transformer [72] has become a research hotspot in the field of computer vision, and multi-scale fusion of visual Transformer is also a relatively new processing solution.

B. EVALUATION METRICS

Anchor-based target detection methods have performed very well for medium and large targets, and the performance of small target detection has gradually improved. Due to the sensitivity of anchors to small targets, they may cause slow convergence or even convergence difficulties during training. Although there are now methods to address such problems and achieve good results, the effective way to eliminate the disadvantages of anchors for small targets is to remove anchors and use anchor-free methods. Some existing research has also proven that anchor-free methods can achieve the same effect as anchor-based methods. Anchor-free detectors determine the location of the target through point priors, which are more suitable for detecting small targets compared to anchor-based detectors. Small targets detection requires higher localization accuracy than large targets, so suitable evaluation metrics can greatly improve the location accuracy. Including the center distance between the predicted values and the ground truth values in the evaluation metrics can improve the accuracy of small target localization.

C. SUPER-RESOLUTION

Super-resolution is a popular direction in the field of computer vision, and there is still a lot of room for development in small target detection. Since the super-resolution reconstruction process is relatively independent of target detection, it limits the integration of the two. Small targets have less feature information, while super-resolution methods can effectively solve this problem. Therefore, using super-resolution reconstruction to enrich the details of small targets, and then converting small target detection problems into medium and large target detection can improve the detection accuracy.

D. THE OPTIMIZATION OF YOLO

Although the YOLO family has made significant progress after years of development, further research is needed to solve more practical problems, such as rotated bounding boxes, 3D targets, few samples, aerial scenes, and how to optimize and deploy the researched algorithm with TensorRT.

In addition, although the YOLO series are the leaders in speed-accuracy balance in the field of target detection, their main work is aimed at computer terminals. Currently, edge computing has become an important trend in the development of artificial intelligence (AI). How to make YOLO lighter and faster for embedded AI computing devices such as Nvidia Jetson TX2, Nano, and Raspberry Pi is a worthy question to ponder.

VI. CONCLUSION

In general, our improvements are mainly based on YOLOv5s. In terms of the Backbone, the proposed multi-branching blended attention block based on CSPNet, i.e., CMBAB, is introduced at the end. In terms of the feature fusion

network, the small target detection branch is added, and proposed MBAB modules are inserted after each layer of the feature pyramid fusion and CMBAB & MBAB modules are inserted after each layer of down-sampling. Finally, the proposed small target detection method MBAB-YOLO is experimentally evaluated on the open source dataset, and the results show that MBAB-YOLO has excellent detection performance, with high accuracy and fast speed, which can meet real-time detection needs.

REFERENCES

[1] X. Li, Z. Xie, T. Lai, F. Zhao, H. Xu, and R. Chen, "NAS-WFPN: Neural architecture search weighted feature pyramid networks for object detection," in *Proc. Int. Workshops Secur., Privacy Anonymity Comput., Commun. Storage*, 2020, pp. 384–394.

[2] F. Xu, L. Duan, and Y. Qiao, "BPN: Bidirectional path network for instance segmentation," in *Proc. CAAI Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, 2021, pp. 55–66.

[3] B. Chen, G. Ghiasi, H. Liu, T. Lin, D. Kalenichenko, H. Adam, and Q. V. Le, "MnasFPN: Learning latency-aware pyramid architecture for object detection on mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13604–13613.

[4] Y. Liu, F. Yang, and P. Hu, "Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks," *IEEE Access*, vol. 8, pp. 145740–145750, 2020.

[5] S. Jo, Y. Ju, K. Cho, S. Kang, M. Ryu, and J. Song, "SATI: Scalable and traffic efficient data dissemination infrastructure for sensor-based distributed information services," Dept. Comput. Sci., Kaist, Daejeon, South Korea, Tech. Rep. CS/TR-2006-265, 2022.

[6] L. Cai, H. Li, W. Dong, and H. Fang, "Micro-expression recognition using 3D DenseNet fused squeeze-and-excitation networks," *Appl. Soft Comput.*, vol. 119, Apr. 2022, Art. no. 108594.

[7] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, and H. Zhang, "Automated ECG classification using a non-local convolutional block attention module," *Comput. Methods Programs Biomed.*, vol. 203, May 2021, Art. no. 106006.

[8] S. Chen and H. Zhang, "Person re-identification based on frequency channel attention networks under the surveillance scenario," *J. Phys., Conf.*, vol. 1966, no. 1, Jul. 2021, Art. no. 012025.

[9] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[10] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 437–446.

[11] B. Li, "Equalized focal loss for dense long-tailed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6990–6999.

[12] Y. Li, W. Zhang, Y. Cai, Z. Li, and X. Jiang, "SNIPER based multi-target and multi-scale aerial image processing method," *J. Phys., Conf.*, vol. 1659, no. 1, Oct. 2020, Art. no. 012003.

[13] J. Liu, D. Li, R. Zheng, L. Tian, and Y. Shan, "RankDetNet: Delving into ranking constraints for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 264–273.

[14] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[15] J. Ma and B. Chen, "Dual refinement feature pyramid networks for object detection," Tech. Rep., 2020.

[16] S. Yang and N. Kumar, "Path aggregation network for multi-organ nuclei segmentation rank-8, team name," Tech. Rep., 2020.

[17] C. Ping-Yang, J. Hsieh, M. Gochoo, and Y. Chen, "Light-weight mixed stage partial network for surveillance object detection with background data augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3333–3337.

[18] S. Singh, "A novel mask R-CNN model to segment heterogeneous brain tumors through image subtraction," 2022, *arXiv:2204.01201*.

[19] F. Nan, W. Jing, F. Tian, J. Zhang, K.-M. Chao, Z. Hong, and Q. Zheng, "Feature super-resolution based facial expression recognition for multi-scale low-resolution images," *Knowl.-Based Syst.*, vol. 236, Jan. 2022, Art. no. 107678.

[20] R. Zhang, Y. Zeng, and X. Jin, "Optimization of small object detection based on generative adversarial networks," in *Proc. ES Web Conf.*, 2021, p. 03062.

[21] H. Wang, J. Wang, K. Bai, and Y. Sun, "Centered multi-task generative adversarial network for small object detection," *Sensors*, vol. 21, no. 15, p. 5194, Jul. 2021.

[22] 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.

[23] J. Rabbi, "Tiny object detection in remote sensing images: End-to-end super-resolution and object detection with deep learning," Tech. Rep., 2020.

[24] L. Chen, C. Liu, F. Chang, S. Li, and Z. Nie, "Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery," *Neurocomputing*, vol. 451, pp. 67–80, Sep. 2021.

[25] Y. Wang, S. M. A. Bashir, M. Khan, Q. Ullah, R. Wang, Y. Song, Z. Guo, and Y. Niu, "Remote sensing image super-resolution and object detection: Benchmark and state of the art," *Exp. Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116793.

[26] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network," *Remote Sens.*, vol. 12, no. 9, p. 1432, May 2020.

[27] K. Zhang, D. Tao, X. Gao, X. Li, and Z. Xiong, "Learning multiple linear mappings for efficient single image super-resolution," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 846–861, Mar. 2015.

[28] C.-H. Fu, H. Chen, H. Zhang, and Y.-L. Chan, "Single image super resolution based on sparse representation and adaptive dictionary selection," in *Proc. 19th Int. Conf. Digit. Signal Process.*, Aug. 2014, pp. 4311–4322.

[29] J. Yang, L. Xiao, and Y. Q. Zhao, "Hybrid local and nonlocal 3-D attentive CNN for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1274–1278, Jul. 2020.

[30] F. Chen, C. Zhu, Z. Shen, H. Zhang, and M. Savvides, "NCMS: Towards accurate anchor free object detection through $\ell 2$ norm calibration and multi-feature selection," *Comput. Vis. Image Understand.*, vol. 200, Nov. 2020, Art. no. 103050.

[31] W. Liu, I. Hasan, and S. Liao, "Center and scale prediction: A box-free approach for pedestrian and face detection," Tech. Rep., 2019.

[32] E. Spyrou and Y. Avrithis, *High-Level Concept Detection in Video Using a Region Thesaurus*. 2007.

[33] M. Zhu and Y. Wu, "A parallel convolutional neural network for pedestrian detection," *Electronics*, vol. 9, no. 9, p. 1478, Sep. 2020.

[34] M. G. Christel and A. G. Hauptmann, *The Use and Utility of High-Level Semantic Features in Video Retrieval*. Berlin, Germany: Springer, 2005.

[35] C. Zhu and Y. Peng, "Discriminative latent semantic feature learning for pedestrian detection," *Neurocomputing*, vol. 238, pp. 126–138, May 2017.

[36] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Frontiers Neurorobotics*, vol. 12, Oct. 2018.

[37] A. F. Smeaton, P. Over, and W. Kraaij, *High-Level Feature Detection from Video in TRECVid: A 5-Year Retrospective of Achievements*. Cham, Switzerland: Springer, 2009.

[38] H.-S. Min, J. Y. Choi, W. De Neve, and Y. M. Ro, "Bimodal fusion of low-level visual features and high-level semantic features for near-duplicate video clip detection," *Signal Process., Image Commun.*, vol. 26, no. 10, pp. 612–627, Nov. 2011.

[39] X. Ye, F. Xiong, J. Lu, H. Zhao, and J. Zhou, "M^2-Net: A multi-scale multi-level feature enhanced network for object detection in optical remote sensing images," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2020, pp. 1–8.

[40] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.

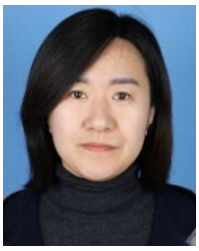
[41] A. Roo, "Towards more robust advice: Message flow analysis for composition filters and its application," Tech. Rep., 2007.

[42] Y. Huang, "Towards more efficient and flexible face image deblurring using robust salient face landmark detection," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 1–20, 2015.
[43] Z. Fang, J. Ren, H. Sun, S. Marshall, J. Han, and H. Zhao, "SAFDet: A semi-anchor-free detector for effective detection of oriented objects in aerial images," *Remote Sens.*, vol. 12, no. 19, p. 3225, Oct. 2020.
[44] C. Sun, "ReAFFPN: Rotation-equivariant attention feature fusion pyramid networks for aerial object detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 3055–3058.
[45] S. Ali, A. Siddique, H. F. Ates, and B. K. Güntürk, "Improved YOLOv4 for aerial object detection," in *Proc. 29th Signal Process. Commun. Appl. Conf. (SIU)*, Jun. 2021, pp. 1–4.
[46] D. Liang, Q. Geng, Z. Wei, D. A. Vorontsov, E. L. Kim, M. Wei, and H. Zhou, "Anchor retouching via model interaction for robust object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[47] X. He, S. Ma, L. He, L. Ru, and C. Wang, "Multi-sector oriented object detector for accurate localization in optical remote sensing images," *Remote Sens.*, vol. 13, no. 10, p. 1921, May 2021.
[48] F. Yang, W. Li, H. Hu, W. Li, and P. Wang, "Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images," *Sensors*, vol. 20, no. 6, p. 1686, Mar. 2020.
[49] J.-J. Ponciano, M. Roetner, A. Reiterer, and F. Boochs, "Object semantic segmentation in point clouds—Comparison of a deep learning and a knowledge-based method," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 4, p. 256, Apr. 2021.
[50] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "Multi-modal 3D object detection in autonomous driving: A survey," *Int. J. Comput. Vis.*, May 2023.
[51] A. Pacheco and R. A. Krohling, "Recent advances in deep learning applied to skin cancer detection," 2019, *arXiv:1912.03280*.
[52] J. Redmon, "You only look once: Unified, real-time object detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[53] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[54] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[55] Q. Li, J. Ning, J. Yuan, and L. Xiao, "A depthwise separable dense convolutional network with convolution block attention module for COVID-19 diagnosis on CT scans," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104837.
[56] M. Everingham, *The 2005 PASCAL Visual Object Classes Challenge* (Lecture Notes in Computer Science), 2006.
[57] X. Zou, Z. Wu, W. Zhou, and J. Huang, "YOLOX-PAI: An improved YOLOX, stronger and faster than YOLOv6," 2022, *arXiv:2208.13040*.
[58] A. R. Siyal, Z. Bhutto, S. Muhammad, A. Iqbal, F. Mehmood, A. Hussain, and S. Ahmed, "Still image-based human activity recognition with deep representations and residual learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 1–7, 2020.
[59] S. K. Yongshin, "Proposing the development of a one to one learning environment to enhance students learning capability," *Korean J. Learn. Sci.*, vol. 8, no. 2, pp. 207–218, 2014.
[60] M. L. Mekhalfi, C. Nicoló, Y. Bazi, M. M. A. Rahhal, N. A. Alsharif, and E. A. Maghayreh, "Contrasting YOLOv5, transformer, and EfficientDet detectors for crop circle detection in desert," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[61] C. Zhang, W. Jiang, and Q. Zhao, "Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision," *Remote Sens.*, vol. 13, no. 6, p. 1176, Mar. 2021.
[62] D. Li, X. Hu, S. Wang, C. Zhang, R. Zhou, and H. Zhou, "Hyperspectral images ground object recognition based on split attention," in *Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Mar. 2021, pp. 324–330.
[63] Y. Liu, G. Zhang, H. Wang, W. Zhao, M. Zhang, and H. Qin, "An efficient super-resolution network based on aggregated residual transformations," *Electronics*, vol. 8, no. 3, p. 339, Mar. 2019.
[64] T. Alipour-Fard, M. E. Paoletti, J. M. Haut, H. Arefi, J. Plaza, and A. Plaza, "Multibranch selective kernel networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 1089–1093, Jun. 2021.
[65] L. Xie and C. Huang, "A residual network of water scene recognition based on optimized inception module and convolutional block attention module," in *Proc. 6th Int. Conf. Syst. Informat. (ICSAI)*, Nov. 2019, pp. 1174–1178.
[66] N. Jiang, X. Yu, X. Peng, Y. Gong, and Z. Han, "SM+: Refined scale match for tiny person detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1815–1819.
[67] W. Jian and L. Lang, "Face mask detection based on transfer learning and PP-YOLO," in *Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Mar. 2021, pp. 106–109.
[68] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "TransVOD: End-to-end video object detection with spatial–temporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7853–7869, Jun. 2023.
[69] M. Durve, S. Orsini, A. Tiribocchi, A. Montessori, J.-M. Tucny, M. Lauricella, A. Camposeo, D. Pisignano, and S. Succi, "Benchmarking YOLOv5 and YOLOv7 models with DeepSORT for droplet tracking applications," *Eur. Phys. J. E*, vol. 46, no. 5, pp. 1–13, May 2023.
[70] Z. Dai, "Uncertainty-aware accurate insulator fault detection based on an improved YOLOX model," *Energy Rep.*, vol. 8, pp. 12809–12821, Nov. 2022.
[71] R. R. Selvaraju, "Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization," Tech. Rep., 2016.
[72] M. Popescu, N. E. Mastorakis, and L. N. Popescu-Perescu, "New aspects providing transformer models," *Perescu*, to be published.

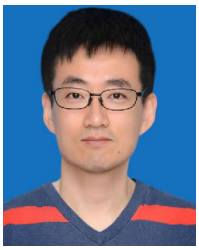
JUN ZHANG was born in 1982. He received the B.S. degree from Northeast Normal University, Changchun, China, in 2003, and the M.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2009. He is currently a Lecturer with the Computer Science Department, Tangshan Normal University, Tangshan, China. His research interests include target detection, target tracking, and deep learning.

YIZHEN MENG received the B.E. degree from the North China University of Science and Technology, Tangshan, China, in 2003, and the M.S. degree from the School of Software Engineering, Tongji University, Shanghai, China, in 2009. She is currently a Lecturer with the Computer Science Department, Tangshan Normal University, Tangshan. Her current research interests include computer vision and pattern recognition.

XIAOHUI YU was born in 1970. She received the bachelor's degree in automation from the Hebei University of Science and Technology, in 1992, and the master's degree from Beijing Jiaotong University, in 2004. She is currently a Lecturer with the Computer Science Department, Tangshan Normal University, Tangshan, China. She is also an associate professor. Her research interest includes neural networks.



HONGJING BI was born in November 1983. She received the Master of Engineering degree in computer applications technology from the Inner Mongolia University of Science and Technology, China, in 2010. She is currently a Lecturer with the Department of Computer Science, Tangshan Normal University. Her research interests include privacy protection, data mining, and time series forecasting.



ZHIPENG CHEN received the Ph.D. degree in signal and information processing from Beijing Jiaotong University, Beijing, China, in 2019. He is currently an Associate Professor with Tangshan Normal University, China. His research interests include multimedia signal processing, digital forensics, and data hiding.

HUAFENG LI was born in 1979. He received the M.S. degree from the Civil Aviation University of China, Tianjin, China. He is currently a Lecturer with the Computer Science Department, Tangshan Normal University, Tangshan, China. His research interests include network security, social engineering, and evolutionary algorithm.

RUNTAO YANG was born in Qinhuangdao, China, in 1982. He received the Ph.D. degree in measurement technology and instrument from the Hefei University of Technology, in 2020. He is currently a Lecturer with the Department of Computer Science, Tangshan Normal University, China. His research interests include optical fiber sensors, fiber Bragg grating inscription, and fiber laser.

JINGJUN TIAN was born in 1968. He received the master's degree in computer software and theory from Northwest University, Xi'an, China, in 2002. He is currently an Associate Professor with the Department of Computer Science, Tangshan Normal University, Hebei, China. His research interests include artificial intelligence and distributed systems.