IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 August 2024, accepted 2 October 2024, date of publication 14 October 2024, date of current version 1 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3479320

RESEARCH ARTICLE

YOLO-Remote: An Object Detection Algorithm for Remote Sensing Targets

KAIZHE FAN[1], QIAN LI[2], QUANJUN LI[1], GUANGQI ZHONG[1], YUE CHU[1], ZHEN LE[1], YELING XU[1], AND JIANFENG LI[1]

[1]School of Advanced Manufacturing, Guangdong University of Technology, Jieyang 522000, China
[2]School of Electronics and Information Engineering, Wuyi University, Jiangmen 510006, China

Corresponding author: Jianfeng Li (li.jianfeng@gdut.edu.cn)

This work was supported in part by the Special Projects in Key Fields of Colleges and Universities of Guangdong Province (New Generation Information Technology) under Grant 2021ZDZX1113.

ABSTRACT Unmanned Aerial Vehicles (UAVs) are indispensable in promoting the development of remote sensing technology. Nevertheless, the tasks of object recognition in remote sensing images based on UAV platforms face major difficulties and challenges due to the complex and variable background environments and the high-density distribution of objects. This paper proposes an object detection algorithm for UAV remote sensing images—YOLO-Remote, which aims to improve detection accuracy by enhancing YOLOv8. This algorithm innovatively integrates the SaElayer module to enhance the focus on remote sensing targets and improve network efficiency. Additionally, it introduces the Efficient-SPPF structure, which effectively expands the network's receptive field and promotes deep learning capabilities. To address sample imbalance and improve bounding box localization and classification performance, the study also designs the Focaler-MDPIOU strategy. With these comprehensive optimizations, YOLO-Remote achieves significant progress in network architecture. Experiments were conducted on the NWPU VHR10 and RSOD datasets, and the experimental results show that compared to the base model YOLOv8n, the improved model's average precision increased by 2.7% and 3.2% respectively, demonstrating its superiority in the field of object detection for UAV remote sensing images.The code is available at https://github.com/QuincyQAQ/Yolo-Remotehttps://github.com/QuincyQAQ/Yolo-Remote.

INDEX TERMS Object detection, YOLOv8, SPPF, remote sensing images.

I. INTRODUCTION

In recent years, the demand for remote sensing targets has been growing across various fields, covering a wide range of applications. For example, in battlefield monitoring [1], disaster response [2], environmental research [3], power maintenance [4], and surveillance and inspection [5], drones have shown outstanding performance in remote sensing tasks, significantly improving operational efficiency. Compared to traditional satellite remote sensing technology, current remote sensing images have made great strides in clarity and accuracy. Nonetheless, challenges remain in capturing distant targets, small objects, heavily obscured items, and weak feature recognition.

In the field of drone-based remote sensing, the challenges of target recognition mainly stem from variations in image scale, uneven and dense distribution of objects, especially the frequent occurrence of small-sized targets [6], [7]. Unlike natural images taken from the ground, the wide-field view captured by drones provides more comprehensive visual information but also complicates scene composition and diversifies target categories, inadvertently increasing background noise in the target detection process. Additionally, aerial observation at a medium scale, due to long-distance shooting, background obstacles, and changing lighting conditions, often makes target identification more difficult. In practice, tasks such as fine-grained classification of vehicle types frequently arise, and the existence of such highly similar targets undoubtedly raises higher requirements for the detection model's accurate differentiation capability.

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

Traditional target detection techniques revolve around three core steps: feature extraction, classification mechanism, and region selection [8]. This process begins with searching for potential target regions within candidate images, followed by extracting features and performing classification. Given that targets can appear in any corner of the image and their sizes and aspect ratios are uncertain, it is necessary to use a multi-scale sliding window strategy to traverse the image to locate potential targets. Although this method can roughly mark the target's location, its high time complexity, window redundancy, and poor region matching significantly hinder the efficiency and quality of feature extraction. Especially when the target's aspect ratio changes significantly, even a full image scan may fail to capture well-matched feature regions, highlighting the severity of time consumption issues. In the feature extraction stage, common techniques include local binary patterns, scale-invariant feature descriptors, and histograms of oriented gradients. However, the variability in target shapes, complex lighting conditions, and diverse background environments greatly challenge the robustness of features, making it difficult to consistently maintain high efficiency [9].

Overall, traditional detection methods have highly variable effectiveness, are easily influenced by external conditions, and have significant limitations in practical applications.

With the accelerated and deepened progress of technology, visual target detection occupies a crucial position in practical deployments. In recent years, numerous tech startups like Sense Time and Megvii have emerged in this field [10]. At the same time, the importance of computer vision technology in the autonomous driving industry is increasingly evident, with pioneering companies like Tesla leading the innovation in visual perception technology for autonomous driving. Despite a series of advances in drone visual detection, challenges remain, mainly due to two aspects: first, the fundamental differences between drone-collected images and natural scene images increase the difficulty of precise target recognition; second, general detection algorithms are inadequate in handling the variability of target sizes in remote sensing images, resulting in unsatisfactory performance [11].

To overcome these challenges, this study designs a model specifically for remote sensing target detection based on the advanced YOLOv8 framework—YOLO-Remote. Experiments on the NWPU VHR10 and RSOD public datasets demonstrate its outstanding detection accuracy. Specific improvements include: first, integrating an innovative Sae-Layer component in the backbone network, significantly enhancing the model's detection accuracy; second, adopting an optimized spatial pyramid pooling and fusion (SPPF) module, effectively broadening the network's receptive field and strengthening feature extraction efficiency; finally, introducing an original Focaler-IOU strategy, effectively addressing the sample imbalance problem in target detection for remote sensing images.

II. RELATED WORK

A. TARGETED DETECTION

Target detection technology, as a crucial branch of computer vision, plays a decisive role in the performance of various visual tasks and applications, thus becoming a focal point across industries. In academic research circles, it is a core topic within computer vision publications, with a significant number of papers discussing target detection published annually. According to statistics, over the past decade, the number of related papers recorded in the Google Scholar database has exceeded 15,000. In industrial practice, many tech giants such as Google, Sense Time, Megvii, Facebook, Huawei, and Baidu have made substantial investments in this field, assembling research and development teams to explore it in depth. Additionally, from a policy perspective, target detection is regarded as a crucial component of the artificial intelligence technology matrix, with countries worldwide actively promoting research and application expansion in this area.

In the early days, target detection algorithms mainly relied on manually designed features combined with simple classifier operations, with Adaptive Boosting (AdaBoost) [12] being a typical example. During this period, a series of classic target feature description algorithms emerged, including Haar features and Histogram of Oriented Gradients (HOG) features. Since 2012, however, with the rapid advancement of deep learning technology, significant improvements in computing power, and the emergence of large-scale open datasets and evaluation standards, a series of milestone research achievements such as Region-based Convolutional Neural Networks (R-CNNs) [13], SSD [14], You Only Look Once (YOLO) [15], and Detection DETR [16] have successively emerged. Compared to previous manual feature construction methods, deep learning technology has greatly simplified the feature design process, achieving automatic feature learning and integrating feature extraction and classifier training within the same framework, thereby driving unprecedented rapid development in this field.

Among the schools of target detection technology, single-stage detection models divide the image into multiple cells, each responsible for determining the presence of objects and their types and positions, with YOLO and SSD being prime examples. In contrast, two-stage detection methods execute the task in two steps: first, generating candidate boxes with a high likelihood of containing targets, and then in the second stage, performing detailed classification and precise localization of these boxes, with Faster R-CNN [17] being an outstanding representative of this approach. While two-stage methods are slightly inferior in real-time performance, they are renowned for their higher detection accuracy and excellent performance across multiple datasets.

B. UNMANNED AERIAL VEHICLE TARGET DETECTION

Multi-target recognition from a drone perspective introduces several challenges, such as the increased number of small

objects, feature-poor information contained in a single viewpoint, low detection efficiency due to uneven distribution of target types, noise interference encountered during detection, missed detections and false positives caused by size variations, and inference delays. This chapter discusses the improvement strategies scholars have developed from two perspectives to address these issues.

When using a drone perspective for multi-target recognition, single-stage detectors like the YOLO series and SSD are widely used due to their substantial advantages. Many researchers have focused on the specific algorithmic challenges of the drone perspective and have tackled them:

To address the phenomenon of numerous small targets in a wide-open field of view, Liu et al. integrated the Liu Res Unit_2 design into the backbone network of YOLO and the ResNet module, and combined dual ResNet units in the residual block of Darknet. This effectively mitigates the problem of small target omission caused by observational limitations, considering that the limited observational scope reduces probability estimation [18]. Researchers like Saetchnikov introduced the YOLOv4 eff model, which adopts a backbone and neck network structure with quadruple cross-stage partial connections and uses the Swish activation function, setting the letter-box size to 1 to maintain efficient utilization [19]. To overcome object misdetection caused by size variations in drone overhead images, Li et al. designed an SSD variant that combines attention mechanisms and dilated convolutions, using dilated convolutions to replace traditional convolutions and integrating low-level feature maps of small-sized objects with high-level feature maps for processing [20]. Compared to single-stage detection algorithms, two-stage target recognition algorithms exhibit different working principles. Directly applying ground perspective algorithms to drone-shot videos is ineffective and requires specific optimizations based on the characteristics of drone images. Key improvements can be summarized as follows:

Avola constructed a multi-stream architecture to process multi-scale images, adapting to the dense conditions of small targets in sky scenes. This architecture was integrated into Fast R-CNN as the backbone, forming the MS-Faster R-CNN detector to ensure continuous stable detection in drone video sequences [21]. Stadler, on the other hand, utilized Cascade R-CNN as the detector, reducing the default anchor box size to match smaller targets and increasing the total number of predicted targets. To address insufficient feature information from a single viewpoint, Azimi et al. used a joint network to extract visual features and combined a graph convolutional neural network with a long short-term memory network (LSTM) to comprehensively analyze the visual, structural, and time-series features of the targets [22]. To solve the problem of processing speed reduction caused by dispersed targets in the sky environment, Yang incorporated the concept of clustering and proposed the ClusDet system. This system first uses the clustering network CPNet to generate target clustering regions, then employs ScaleNet to evaluate the target sizes within these regions, and subsequently sends these regions to DetecNet for target recognition, thereby reducing the computational burden and ultimately achieving efficient detection [23].

III. PRINCIPLES AND IMPROVEMENTS

A. YOLOv8

In comparison with YOLOv5 and YOLOv7 algorithms, YOLOv8 has achieved significant improvements in shortening training cycles and enhancing recognition accuracy. Additionally, its model weight file occupies only 6MB of space, making it easily deployable to any embedded device. With its rapid and efficient operation, it is well-suited for real-time detection tasks.As the successor to YOLOv5, YOLOv8 inherits and further develops its predecessor, offering models in various sizes including N, S, M, L, and X to accommodate diverse application scenarios. This algorithm not only achieves significant breakthroughs in accuracy but also ensures a smooth training process and broad hardware platform compatibility, enabling flexible deployment.For input processing, YOLOv8 employs innovative data augmentation strategies such as Mosaic technology and adaptive anchor box estimation algorithm. Mosaic technology enhances the diversity of the detection dataset through random scaling, cropping, and layout reorganization of images. Adaptive anchor box calculation optimizes anchor box configuration through precise difference computation and reverse iteration based on the initial anchor box prediction output.On the output side, YOLOv8 revolutionarily replaces the traditional coupled head design with a decoupled head structure, separating classification and regression tasks into two independent branches. This decoupling strategy allows each task to focus more effectively, addressing localization deviations and classification errors in complex scenes. Furthermore, the algorithm incorporates the DFL strategy and implements an anchor-free target detection method, enabling the network to quickly lock onto the target's surrounding area. This results in prediction boxes that closely fit the actual boundaries, thereby enhancing detection accuracy.

B. IMPROVEMENT

1) SaELayer

The SaElayer module [24] is an innovative design, as shown in Figure 1, that skillfully combines the efficient characteristics of the Squeeze-and-Excitation Network (SENet) module with the inter-layer dense communication advantages of DenseNet, aiming to enhance network performance. Furthermore, this module creatively incorporates fully connected layer designs with multi-scale branches that have different width configurations. This strategy greatly enhances the network's ability to capture and integrate global contextual information, providing a more comprehensive and in-depth understanding for target detection in complex remote sensing images. By integrating the SaElayer into

YOLOv8n, the network's attention to critical detection features in remote sensing images is significantly enhanced, allowing it to more keenly identify and focus on target areas. Additionally, it achieves effective network resource management. Specifically, this design optimizes the use of network bandwidth, effectively reducing the computational resources and time costs required during model training, thus accelerating the model convergence process and improving training efficiency.

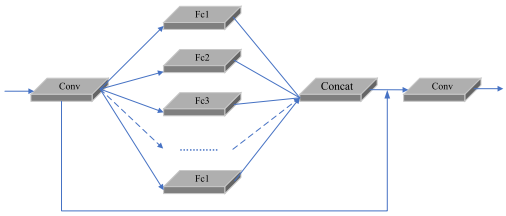


FIGURE 1. Schematic diagram of the SaElayer structure.

2) EFFICIENT-SPPF

YOLOv8 adopts a structure called Spatial Pyramid Pooling Fusion (SPPF), which combines serial and parallel pooling mechanisms to broaden the network's receptive field. However, this approach may show limitations in certain situations due to its fixed-size pooling strategy, which may not fully accommodate the multi-scale requirements inherent in remote sensing image target detection tasks. Additionally, it struggles to capture the fine details and comprehensive contextual information crucial for the resolution of remote sensing targets.

Integrating a broader receptive field into the deep neural network architecture is particularly important for enhancing the model's contextual understanding capabilities, which is critical for accurately performing remote sensing target detection tasks. Expanding the receptive field helps encompass a wider range of surrounding information, reducing misjudgments and improving the model's semantic understanding and feature extraction efficiency, especially when handling complex scenes or detecting small objects. It ensures that each convolutional output contains more information.

Common methods for increasing the receptive field include applying additional convolution and pooling steps to the feature map. However, these operations can lead to the loss of feature information and come with higher computational costs. As a widely used technique in the field of image segmentation, ''dilated convolution'' [25] successfully expands the receptive field while maintaining the original resolution of the feature map, bypassing the downsampling and upsampling steps. This convolution technique introduces a ''dilation rate'' parameter that defines the spacing between pixels when the convolution kernel processes them. The specific differences between standard convolution and dilated convolution are illustrated in Figure 2.

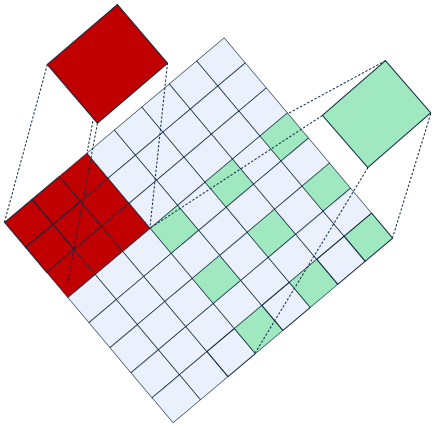


FIGURE 2. The red represents standard convolution (dilation rate = 1, receptive field = 3); the green represents dilated convolution (dilation rate = 2, receptive field = 5).

Therefore, we chose to incorporate dilated convolution technology into the existing SPPF module, naming it Efficient-SPPF. The structural details are shown in Figure 3. The specific improvements are summarized as follows:

1) After the final max-pooling step of the standard SPPF structure, a set of parallel dilated convolution layers are added, using dilation rates of 2, 4, and 8, thereby constructing a diversified receptive field to cover convolution kernels of different scales.
2) A residual connection path is introduced to alleviate the gradient vanishing problem and enhance the model's ability to capture global image features. This residual path consists of an average pooling layer, a single 1×1 convolution layer, and an upsampling operation.
3) Two customized fusion strategies are adopted for different channel numbers of the feature map to optimize information integration.

The Efficient-SPPF design not only deepens the network structure and broadens the receptive field but also ensures the efficiency of the model during deep learning while maintaining the original resolution of the feature map. Through this design, the model can extract image features from multiple dimensions, thereby comprehensively grasping contextual and background information, effectively compensating for the remote sensing target details that might be missed in traditional feature extraction processes, especially for small objects and targets in complex backgrounds. These series of improvements significantly enhance the detection accuracy and generalization ability of our model.

3) Focaler-MDPIoU

a: FOCALER-IoU

When performing remote sensing target recognition tasks, encountering the problem of sample imbalance is a common phenomenon. Samples can be categorized into two types based on the difficulty of detection: easy-to-process samples and challenging samples. From the perspective of target size,

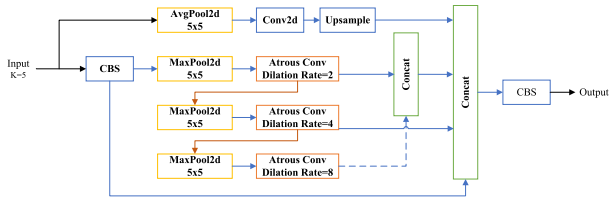


FIGURE 3. The Efficient-SPPF structure diagram.

targets of regular size are usually considered easy-to-process samples, whereas small targets are classified as challenging samples due to the high difficulty in accurate localization. In recognition tasks dominated by easy-to-process samples, focusing on the bounding box regression process for these samples has been proven to effectively enhance detection performance. Conversely, when dealing with tasks primarily composed of challenging samples, it is essential to prioritize and optimize the bounding box regression strategy for these samples to address the detection difficulties.

To accommodate the specific attention to different regression samples in remote sensing target detection tasks, we adopted a linear interval mapping technique to reshape the IoU (Intersection over Union) loss function, aiming to optimize the boundary regression performance. Its mathematical expression is as follows:

$$IoU^{focaler} = \begin{cases} 0, & IoU < d \\ \dfrac{IoU - d}{u - d}, & d \ll IoU \ll u \\ 1, & IoU > u \end{cases} \tag{1}$$

Here, $\text{IoU}^{\text{focaler}}$ represents an improved concept of Focaler-IoU [26], while IoU retains its basic Intersection over Union meaning. Both are set to operate within the range [0,1],specifically with values in the range $[d, u]$. By finely tuning the parameters d and u, we can guide IoUfocaler to give varying degrees of importance to different types of regression instances. The corresponding loss function is described as follows:

$$L_{Focaler-IoU} = 1 - IoU^{focaler} \tag{2}$$

In the process of object detection, BoundingBox Regression (BBR) is a crucial component for achieving precise object localization. Mainstream advanced detection models, including Faster R-CNN, DETR, and the YOLO family, all adopt BBR strategies to accurately pinpoint object locations. The Yolov7-tiny model employs a Comprehensive IoU Loss (CloU) [27],which is an enhanced IoU-based loss function, as the core metric for optimizing its localization performance Compared to the loss functions used in previous YOLO series, the CloU loss function integrates the overlap area between the predicted and ground truth boxes, the distance between their center points, and their aspect ratio information. This integration significantly enhances the precision of bounding box convergence, while also making the regression process smoother and more stable. The mathematical formulation of CIoU is defined as follows:

$$L_{\text{CIoU}} = 1 - IoU + \frac{\rho^2(b, b_{gt})}{c^2} + \alpha v \tag{3}$$

b: MPDIoU

Facing this challenge, Ma and colleagues [28] introduced an innovative loss function on top of the existing CloU loss function—Minimum Points Distance IoU (MPDIoU). The uniqueness of this method lies in its use of the vertex coordinates (top-left and bottom-right) of the predicted and ground truth boxes to comprehensively analyze the presence or absence of overlapping regions, the spatial displacement between the box centers, and the actual size deviations of the box dimensions. This approach optimizes and simplifies the loss calculation process. Specifically, labeling the ground truth box as B_{gt} and the predicted box as B_{prd}, the mathematical formulation of MPDIoU can be reconstructed as follows:

$$\begin{aligned} \text{MPDIoU} &= \frac{B_{gt} \cap B_{prd}}{B_{gt} \cup B_{prd}} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \\ L_{MPDIoU} &= 1 - MPDIoU \end{aligned} \tag{4}$$

In expression 4, w and h represent the horizontal and vertical dimensions of the input image, respectively. d_1 measures the straight-line distance between the top-left vertices of the ground truth box and the predicted box, while d_2 measures the same distance for the bottom-right vertices of these two boxes. IoU is defined as the ratio of the intersection area to the union area of the groumd truth box B_{gt} and the predicted box B_{prd}. In the example shown in Figure 4, assume the dashed box represents the image under analysis, with the pentagon marking the target location within the image to be identified. The solid box on the left represents the actual annotation box B_{gt}, while the solid box on the right is the model's predicted box B_{prd} Specifically, assume the top-left coordinates of Bgt are $(x_1^{\text{gt}}, y_1^{\text{gt}})$. and the bottom-right coordinates are $(x_2^{\text{gt}}, y_2^{\text{gt}})$., while the corresponding top-left coordinates of Bprd are $(x_1^{\text{prd}}, y_1^{\text{prd}})$. and the bottom-right coordinates are $(x_2^{\text{prd}}, y_2^{\text{prd}})$. Then, the conversion formulas for each parameter in the MPDIoU formula can be further detailed as follows:

$$\begin{aligned} d_1^2 &= \left(x_1^{prd} - x_1^{gt}\right)^2 + \left(y_1^{prd} - y_1^{gt}\right)^2 \\ d_2^2 &= \left(x_2^{prd} - x_2^{gt}\right)^2 + \left(y_2^{prd} - y_2^{gt}\right)^2 \\ w_{gt} &= x_2^{gt} - x_1^{gt}, h_{gt} = y_2^{gt} - y_1^{gt} \\ w_{prd} &= x_2^{prd} - x_1^{prd}, h_{prd} = y_2^{prd} - y_1^{prd} \end{aligned} \tag{5}$$

The horizontal and vertical extents of the ground truth box are represented by w_{gt} and h_{gt}, respectively, while the corresponding dimensions of the predicted box are given by w_{prd}. and h_{prd}.. Through further mathematical transformations, this fundamental data can reveal key parameters such as the area of overlap or non-overlap between the predicted and ground truth boxes, the differences in their center coordinates, and the width and height differences. All these components

essentially derive from the basic information set of the top-left and bottom-right vertex coordinates of the ground truth and predicted boxes. This indicates that the MPDIoU loss function deeply exploits the geometric properties of the bounding boxes, while achieving effective simplification and optimization in computation.

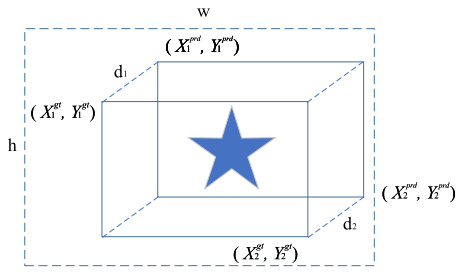


FIGURE 4. Geometric illustration of MPDIoU.

As illustrated in Figure 4, the structure diagram of MPDIoU ensures that when the width and height of the predicted box and the ground truth box maintain a linear proportion, the loss value is lower when the predicted box falls within the ground truth box compared to when it falls outside. This mechanism effectively distinguishes between the two different states, thereby promoting the accuracy of bounding box regression. During model training, MPDIoU also incorporates parameter tuning within the deep regression model, driving each predicted box to converge towards its corresponding ground truth box, with the ultimate goal of reducing the disparity between the bounding boxes.

c: FOCALER-MDPIoU

This article combines the ideas of Focaler-IoU and MPDIOU to propose a new loss function called Focaler-MDPIOU. It aims to alleviate the problem of sample imbalance in remote sensing targets, while improving the localization accuracy and classification performance of detection boxes. The specific formula is as follows:

$$L_{Focaler-MPDIOU} = L_{MPDIOU} + IoU - IoU^{Focaler} \quad (6)$$

C. YOLO-REMOTE ALGORITHM NETWORK

We apply the above improvements to YOLOv8n. As shown in Figure 5, the YOLO-Remote algorithm network diagram, specifically, we use the Efficient-SPPF structural layer to replace the original SPPF in YOLOv8n, thereby expanding the network's receptive field. Additionally, we add the SaElayer after the Efficient-SPPF structural layer to focus more on the feature information of remote sensing targets. We also use Focaler-MPDIOU to alleviate the problem of sample imbalance in remote sensing targets.

IV. ANALYSIS OF EXPERIMENTAL RESULTS

A. DATASETS

To verify the effectiveness of the YOLO-Remote network, we conducted experiments using the NWPU VHR-10 dataset, which contains 3,651 objects across 10 classes: airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. Additionally, we used the RSOD dataset for further validation. This dataset includes four classes of objects: airplane, playground, overpass, and oil tank, with a total of 4993 airplanes in 446 images, 191 playgrounds in 189 images, 180 overpasses in 176 images, and 1586 oil tanks in 165 images. The training set of the RSOD dataset consists of 454 images, the test set consists of 99 images, and the validation set consists of 97 images.

B. EXPERIMENTAL STEPS

All experiments in this paper were conducted using deep learning techniques. To ensure the reliability and consistency of the experimental results, all experiments were performed in a unified environment, and no pre-trained models were used. This means all models were trained from scratch. The experimental configuration, as shown in Table 3.5, mainly includes setting the input image size to 640 × 640 pixels, batch size to 32, training epochs to 300, and initial learning rate to 0.01. We chose SGD as the optimizer, with a momentum parameter of 0.937, and introduced a weight decay factor of 5e-4 to optimize the training process.

C. EVALUATION INDICATORS

The performance of the model is assessed using a set of metrics comprising precision (P), recall (R), mean average precision (mAP), and average precision (AP) per class. AP acts as an indicator for the detection accuracy of individual classes, whereas mAP aggregates the AP scores from all classes and divides them by the total class count to provide an overall performance measure. Specifically, in this research, mAP0.5 denotes the mean average precision at an intersection over union (IoU) threshold of 0.5, which quantifies how well the forecasted bounding boxes align with the ground truth ones.

$$\begin{aligned} P &= TP/(TP+FP) \\ R &= TP/(TP+FN) \\ AP &= \int_0^1 P(R)dR \\ mAP &= \frac{1}{N}\int_0^1 P(R)dR \end{aligned} \quad (7)$$

When assessing the model's efficacy, true positive (TP) indicate the instances where the model accurately classified positive samples. Conversely, false positive (FP) signify the number of times the model incorrectly labeled negative instances as positive. Additionally, false negatives (FN) account for positive samples that the model failed to recognize, instead categorizing them as negative. These fundamental metrics form the basis for determining precision, recall, and additional key performance parameters.

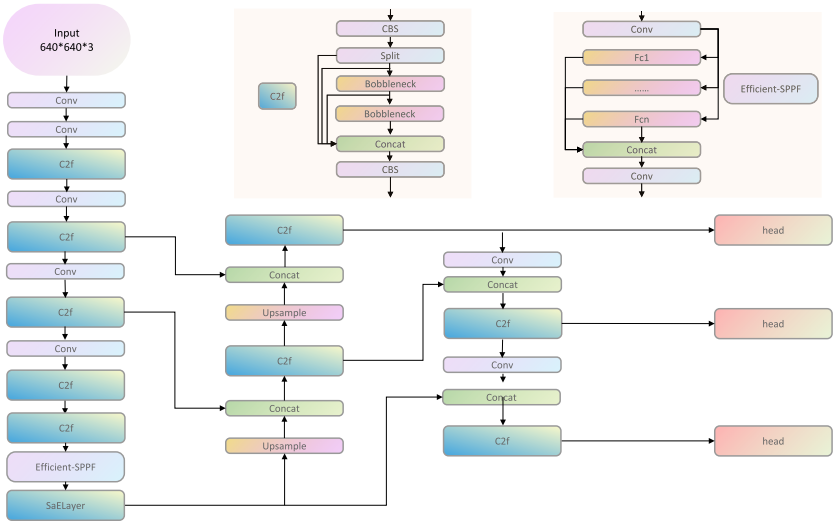


FIGURE 5. YOLO-remote algorithm network.

TABLE 1. Ablation experiments on the NWPU dataset.

SaELayer	Efficient-SPPF	Focaler-MPDIOU	P	R	map	F1	Param	GFlops
			0.898	0.835	0.893	0.86	3.00	8.2
√			0.905	0.861	0.906	0.87	3.07	8.2
	√		0.905	0.861	0.906	0.87	3.07	8.2
		√	0.932	0.848	0.909	0.88	3.00	8.2
	√	√	0.902	0.850	0.910	0.87	3.07	8.2
√	√	√	0.909	0.886	0.920	0.90	3.12	8.3

D. ABLATION EXPERIMENTS

To validate the effectiveness of each improvement module in our proposed method, ablation studies were conducted on the aforementioned dataset. The Sequential Attention Enhancement Layer (SaELayer), the Efficient-SPPF attention module, and the Focal Refinement with Multi-Distance Peak Intersection-over-Union loss (+Focaler-MDPIOU) were sequentially integrated into the baseline model. The experimental outcomes are presented in the table provided.

Table 1 illustrates the results of the ablation experiments conducted on the NWPU VHR-10 dataset. In this study, our approach incrementally incorporated three improvement modules - SaELayer, Efficient-SPPF, and Focaler-MDPIOU - into the YOLOv8n algorithm, observing varying degrees of performance enhancement with each addition. Specifically, the separate integration of SaELayer and Efficient-SPPF respectively led to an increase in accuracy by 0.7%, a rise in mean Average Precision (mAP) by 1.6%, and an augmentation in F1 score by 1.3 points. Notably, the inclusion of Focaler-MDPIOU yielded the most substantial improvements, boosting accuracy by 3.4%, mAP by 1.6%, and the F1 score by 1.7 points. When both Efficient-SPPF and Focaler-MDPIOU were applied concurrently, the mAP saw a 1.7% enhancement compared to the baseline network.

The integration of all three modules, culminating in the YOLO-Remote algorithm, achieved the optimal uplift across all evaluation metrics. This comprised an accuracy improvement of 1.1%, a significant rise in recall by 5.1%, a 2.7% boost in mAP, and a 2.7-point increase in the F1 score. Remarkably, this comprehensive upgrade came with only a marginal increase in both the number of parameters and floating-point operations. Collectively, the outcomes of these ablation experiments robustly substantiate the efficacy of the three introduced modules and the YOLO-Remote algorithm.

Table 2 presents the outcomes of the ablation experiments conducted on the RSOD dataset. Building upon the YOLOv8n algorithm, we individually integrated the SaELayer, Efficient-SPPF, and Focaler-MDPIOU enhancement modules, observing remarkable performance enhancements. Specifically, the introduction of SaELayer alone resulted in a 2.3% increase in mAP and a 1.0% rise in F1 score. Following the integration of Efficient-SPPF, accuracy was enhanced by 1.4%, mAP rose by 0.9%, and the F1 score also saw a 1.0% improvement. Furthermore, the inclusion of Focaler-MDPIOU led to a 1.8% boost in precision, a 1.6% growth in mAP, and a 1.7% increase in the F1 score. When both SaELayer and Efficient-SPPF were employed simultaneously, the mAP witnessed a 1.7% uplift compared to the baseline model.The synergy of all three modules, embodied in the YOLO-Remote algorithm, yielded the optimal enhancements across all evaluation criteria, achieving an mAP of 92.7%, representing a 3.2% increase

TABLE 2. Ablation experiments on the RSOD dataset.

SaELayer	Efficient-SPPF	Focaler-MPDIOU	P	R	map	F1	Param	GFlops
			0.911	0.85	0.895	0.88	3.00	8.2
√			0.881	0.934	0.918	0.90	3.02	8.2
	√		0.925	0.875	0.904	0.89	3.07	8.2
		√	0.903	0.867	0.913	0.88	3.07	8.2
	√	√	0.887	0.892	0.910	0.89	3.12	8.2
√	√	√	0.917	0.893	0.927	0.91	3.12	8.3

TABLE 3. Comparative experiments of different algorithms on the NWPU dataset.

Model	P	R	map	F1	Param	GFlops
Yolov5n	0.913	0.84	0.895	0.87	2.50	7.1
Yolov6n	0.894	0.855	0.903	0.86	4.23	11.8
Yolov8n	0.898	0.835	0.893	0.86	3.02	8.2
RT-DETR	0.764	0.782	0.828	0.77	3.19	103.4
RT-DETR-resnet101	0.856	0.844	0.89	0.77	61.91	191.5
Faster-RCNN-resnet101	0.887	0.885	0.8780	0.78	60.74	223.1
YOLO-Remote	0.917	0.893	0.927	0.89	3.12	8.3

over the baseline model, and an F1 score improvement of 3%. These ablation experiment findings decisively validate the efficacy of the three newly introduced modules and the YOLO-Remote algorithm as a whole.

Table 3 The paper presents the YOLO-Remote algorithm, which is compared with several existing algorithms, showcasing its outstanding performance across multiple key metrics and demonstrating a clear advantage over competing models. Specifically, YOLO-Remote achieved an accuracy rate of 92.7%, surpassing Yolov5n by 0.4% and Yolov6n by 2.3%, highlighting its superior capability in correct classification. Its recall rate is 0.893, which is 4.44% higher than Yolov6, indicating better performance in identifying all positive samples. In terms of mean Average Precision (mAP), YOLO-Remote reached 0.927, outperforming Yolov6 by 2.4%, RT-DETR by 9.9%, and RT-DETR-resnet101 by 12%, demonstrating consistent and stable detection performance across different thresholds. Its F1 score is the highest among all models at 0.89, indicating an optimal balance between precision and recall. Additionally, compared to classical two-stage algorithms, YOLO-Remote exhibits significant advantages across all metrics.

Moreover, YOLO-Remote boasts a model size of 3,122,670 parameters and a computational demand of 8.3 GFlops, reflecting a relatively low complexity. This balanced combination of efficiency and performance makes it well-suited for deployment in resource-constrained environments. In summary, the YOLO-Remote algorithm excels not only in delivering high performance but also in maintaining low resource consumption, thereby showcasing its practicality and efficiency in real-world applications.

From Table 4, it is evident that the YOLO-Remote algorithm model demonstrates exceptional performance across multiple key performance indicators, notably outperforming several comparative models. Specifically, the proposed model achieves an accuracy of 0.917, which represents respective improvements of 3%, 2.3%, 0.6%, and 13.4% over Yolov5n, Yolov6n, Yolov8n, and RT-DETR. Its recall rate reaches 89.3%, marking increases of 4.1%, 3.8%, 4.3%, and 23.6% in comparison to these models. In terms of mean Average Precision (mAP), the proposed model attains 92.7%, surpassing Yolov5, Yolov6, Yolov8n, and RT-DETR by 3.5%, 2.4%, 3.2%, and 19.4%, respectively.

Regarding computational resource consumption, although the YOLO-Remote model has a slightly higher number of parameters and computational load compared to Yolov8n, it is considerably less demanding than the RT-DETR model. This illustrates that the YOLO-Remote model not only possesses substantial performance advantages but also maintains a commendable balance in computational efficiency. Overall, by maintaining a low computational cost, the YOLO-Remote model achieves outstanding performance across various evaluation metrics, thereby exemplifying the superiority of the proposed algorithm.

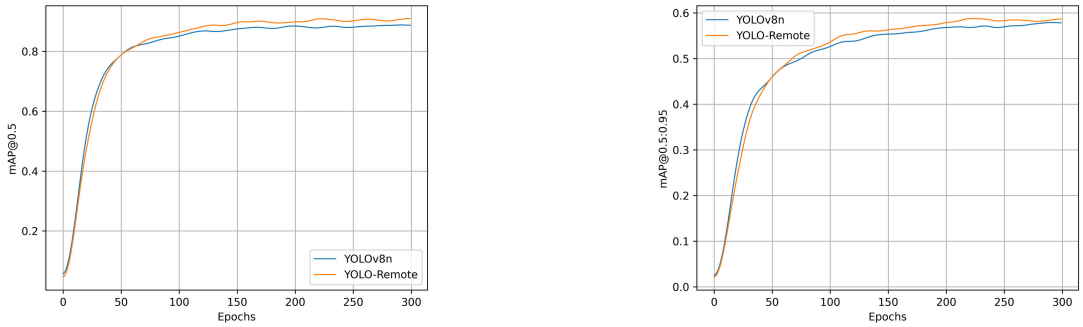
E. ANALYSIS OF EXPERIMENT

Figure 6 illustrates the trends of mAP0.5 and mAP0.5-0.95 for two algorithms on the NMPU dataset under the same number of training epochs. It is evident from the figure that, as the models converge, the improved YOLO-Remote algorithm consistently outperforms the original YOLOv8n algorithm in both mAP0.5 and mAP0.5-0.95. This indicates that YOLO-Remote indeed has a significant advantage in enhancing object detection performance.

Figure 7 presents the evolution trends of the mAP metrics at a 0.5 threshold and within the 0.5 to 0.95 range for two algorithms on the RSOD dataset, after an equal number of training iterations. The figure clearly shows that as the models stabilize through training, the optimized YOLO-Remote algorithm consistently surpasses the basic YOLOv8n algorithm in both mAP metrics. This result strongly demonstrates

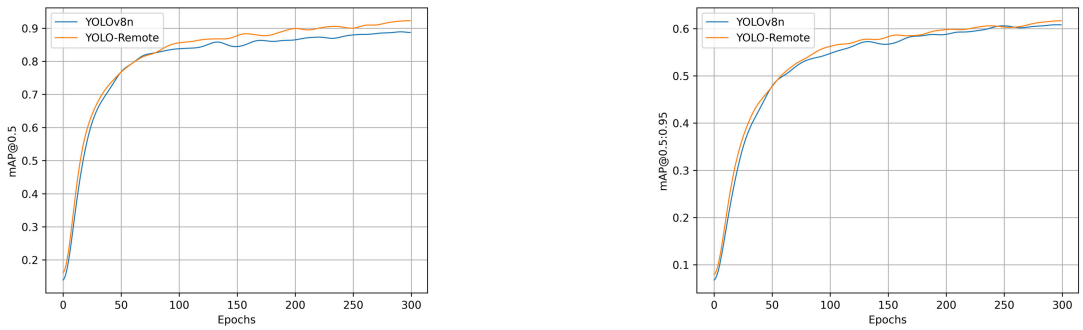
TABLE 4. Comparative experiments of various algorithms on the RSOD dataset.

Model	P	R	map	F1	Param	GFlops
Yolov5n	0.887	0.852	0.892	0.87	2.50	7.1
Yolov6n	0.894	0.855	0.903	0.87	4.23	11.8
Yolov8n	0.911	0.85	0.895	0.88	3.00	8.2
RT-DETR	0.783	0.657	0.733	0.68	31.99	103.4
RT-DETR-resnet101	0.783	0.777	0.811	0.77	61.91	191.5
Faster-RCNN-resnet101	0.889	0.873	0.882	0.835	60.74	223.1
YOLO-Remote	0.917	0.893	0.927	0.90	3.12	8.3



(a) mAP0.5 (b) mAP0.5-0.95

FIGURE 6. Comparative graph of key performance indicators between YOLO-Remote and YOLOv8n on the NWPU dataset.



(a) mAP0.5 (b) mAP50-95

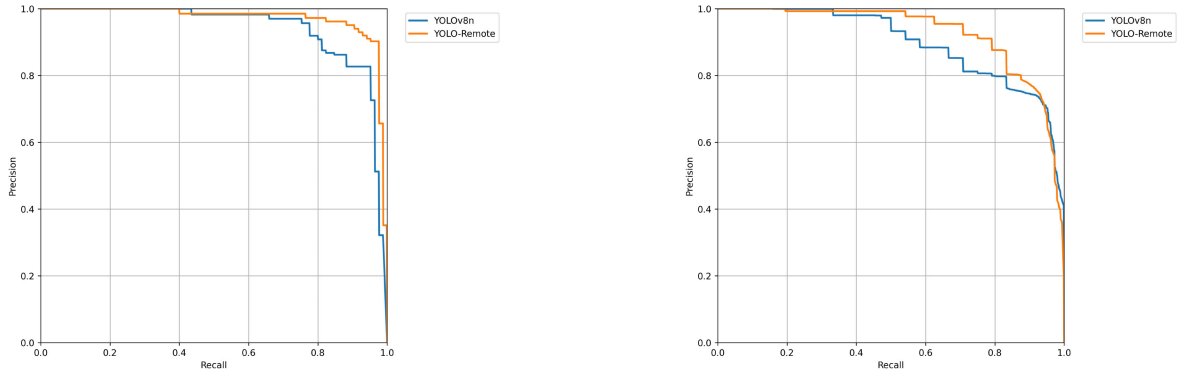
FIGURE 7. Comparative graph of key performance indicators between YOLO-Remote and YOLOv8n on the RSOD dataset.

that YOLO-Remote significantly improves performance in remote sensing image object detection tasks.

Figure 8 shows the Precision-Recall curves of YOLO-Remote and YOLOv8n on the NWPU and RSOD datasets. From the figure, it can be seen that YOLO-Remote (orange curve) maintains higher precision than YOLOv8n (blue curve) at most recall levels. This indicates that YOLO-Remote not only sustains higher precision at high recall rates but also performs more stably across the entire recall range, with a lower false detection rate. Therefore, the improved YOLO-Remote algorithm outperforms YOLOv8n in object detection, offering higher detection precision and stability.

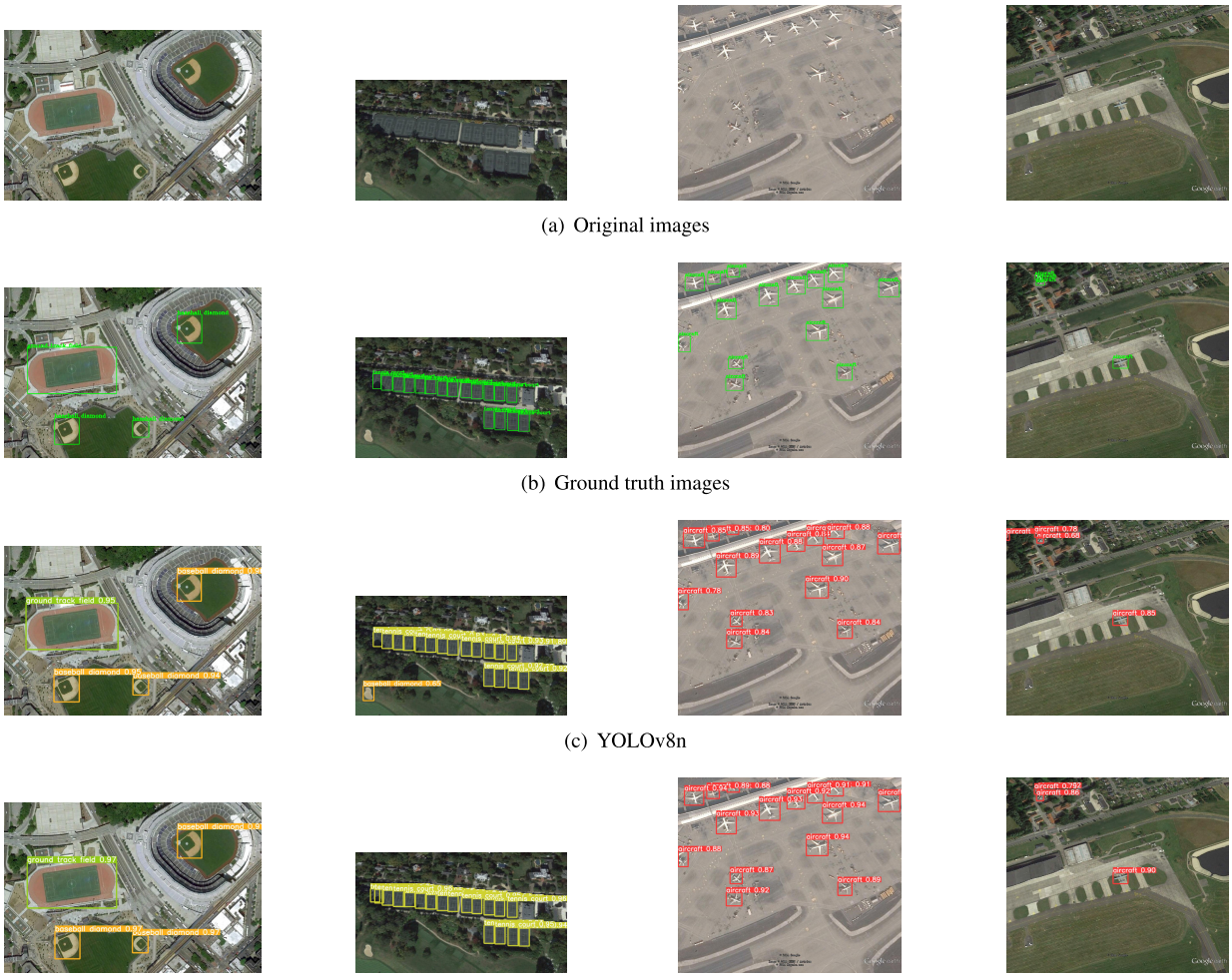
Figure 9 compares the detection performance of the baseline model YOLOv8n and the improved algorithm YOLO-Remote on two remote sensing target datasets. The figure is divided into four rows: the first row shows the original images, the second row shows the ground truth images, the third row shows the detection results of YOLOv8n, and the forth row shows the detection results of the improved YOLO-Remote algorithm. It is clear from the comparison that the improved YOLO-Remote algorithm significantly outperforms the baseline YOLOv8n algorithm in remote sensing target detection.

From the detection results, it can be observed that YOLOv8n tends to miss or falsely detect targets in some complex backgrounds, while YOLO-Remote demonstrates higher accuracy and robustness. Specifically, YOLO-Remote excels in locating target boundaries and capturing target



(a) NWPU dataset (b) RSOD dataset

FIGURE 8. **Comparative graph of key performance indicators between YOLO-Remote and YOLOv8n on the NWPU and RSOD datasets.**



(a) Original images

(b) Ground truth images

(c) YOLOv8n

(d) The improved algorithm YOLO-Remote

FIGURE 9. **Comparative visualization of detection results on two remote sensing target datasets.**

details, indicating that this algorithm has a stronger advantage in handling small targets and complex scenes in remote sensing images. Therefore, the improved YOLO-Remote provides more reliable detection results in practical applications, offering better technical support for remote sensing image analysis.

V. CONCLUSION

In this study, a new algorithm, YOLO-Remote, is proposed. By introducing SaeLayers, Efficient-SPPF with dilated convolution, and Focaler-MPDIOU into the YOLOv8 algorithm, it successfully addresses the challenge of detecting remote sensing targets in complex backgrounds. Experimental validation shows that the improved algorithm in this study performs exceptionally well on both UAV remote sensing image datasets and satellite remote sensing image datasets. Specifically, the mAP on the NWPU dataset increased by 2.7% compared to the baseline model, and the mAP on the RSOD satellite remote sensing image dataset increased by 3.2% compared to the baseline model, demonstrating the strong practicality and effectiveness of the proposed YOLO-Remote algorithm.

ACKNOWLEDGMENT

(Kaizhe Fan, Qian Li, and Quanjun Li are co-first authors.)

REFERENCES

[1] L. Kong, J. Wang, and P. Zhao, "YOLO-G: A lightweight network model for improving the performance of military targets detection," *IEEE Access*, vol. 10, pp. 55546–55564, 2022.
[2] Y. Yang, Z. Miao, H. Zhang, B. Wang, and L. Wu, "Lightweight attention-guided YOLO with level set layer for landslide detection from optical satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3543–3559, 2024.
[3] X. Xu, Y. Liu, L. Lyu, P. Yan, and J. Zhang, "MAD-YOLO: A quantitative detection algorithm for dense small-scale marine benthos," *Ecolog. Informat.*, vol. 75, Jul. 2023, Art. no. 102022.
[4] B. J. Souza, S. F. Stefenon, G. Singh, and R. Z. Freire, "Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV," *Int. J. Electr. Power Energy Syst.*, vol. 148, Jun. 2023, Art. no. 108982.
[5] G. Huanca, J. Abel Ordonez, and C. Helsner Menacho, "Personal protective equipment use inspection, real time surveillance with YOLO," in *Proc. 7th Int. Conf. Mach. Learn. Technol. (ICMLT)*, vol. 13, Mar. 2022, pp. 223–229.
[6] Z. Zhang, "Drone-YOLO: An efficient neural network method for target detection in drone images," *Drones*, vol. 7, no. 8, p. 526, Aug. 2023. [Online]. Available: https://www.mdpi.com/2504-446X/7/8/526
[7] Y. Hui, J. Wang, and B. Li, "DSAA-YOLO: UAV remote sensing small target recognition algorithm for YOLOV7 based on dense residual super-resolution and anchor frame adaptive regression strategy," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 1, Jan. 2024, Art. no. 101863.
[8] Y. Sun, Z. Sun, and W. Chen, "The evolution of object detection methods," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108458. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095219762400616X
[9] Z. Chu, "D-YOLO a robust framework for object detection in adverse weather conditions," 2024, *arXiv:2403.09233*.
[10] B. Li and Y. Gao, "China's competitive advantages in artificial intelligence development," *Sci. Insights*, vol. 43, no. 1, pp. 1003–1007, Jul. 2023.
[11] X. Wu, W. Li, D. Hong, R. Tao, and Q. Du, "Deep learning for UAV-based object detection and tracking: A survey," 2021, *arXiv:2110.12638*.
[12] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Statist. Interface*, vol. 2, no. 3, pp. 349–360, 2009.
[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[14] W. Liu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9905, 2016, pp. 21–37.
[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.
[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[18] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, p. 2238, Apr. 2020.
[19] I. Saetchnikov, V. Skakun, and E. Tcherniavskaia, "Efficient objects tracking from an unmanned aerial vehicle," in *Proc. IEEE 8th Int. Workshop Metrology Aerosp. (MetroAeroSpace)*, Jun. 2021, pp. 221–225.
[20] Z. Li, X. Liu, Y. Zhao, B. Liu, Z. Huang, and R. Hong, "A lightweight multi-scale aggregated model for detecting aerial images captured by UAVs," *J. Vis. Commun. Image Represent.*, vol. 77, May 2021, Art. no. 103058.
[21] D. Avola, L. Cinque, A. Diko, A. Fagioli, G. L. Foresti, A. Mecca, D. Pannone, and C. Piciarelli, "MS-faster R-CNN: Multi-stream backbone for improved faster R-CNN object detection and aerial tracking from UAV images," *Remote Sens.*, vol. 13, no. 9, p. 1670, Apr. 2021.
[22] S. M. Azimi, M. Kraus, R. Bahmanyar, and P. Reinartz, "Multiple pedestrians and vehicles tracking in aerial imagery using a convolutional neural network," *Remote Sens.*, vol. 13, no. 10, p. 1953, May 2021.
[23] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8310–8319.
[24] M. Narayanan, "SENetV2: Aggregated dense layer for channelwise and global representations," 2023, *arXiv:2311.10807*.
[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[26] H. Zhang and S. Zhang, "Focaler-IoU: More focused intersection over union loss," 2024, *arXiv:2401.10525*.
[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000.
[28] S. Ma and Y. Xu, "MPDIoU: A loss for efficient and accurate bounding box regression," 2023, *arXiv:2307.07662*.

KAIZHE FAN was born in Zhuhai, Guangdong, in 2002. He is currently pursuing the degree in electronic and information engineering, with strong hands-on development ability with the School of Advanced Manufacturing, Guangdong University of Technology. His research interests include embedded systems and artifictial intelligence.

QIAN LI is currently pursuing the degree in communication engineering with Wuyi University, with a focus on computer science. Her research interests include object detection in UAV remote sensing images and improving detection accuracy in complex background environments.

QUANJUN LI is currently pursuing the master's degree with the School of Advanced Manufacturing, Guangdong University of Technology. He has published more than two peer-reviewed articles in highquality international journals. His research interests include supplier partner selection and supply chain management. He was a recipient of a number of scholarships and awards.

GUANGQI ZHONG is currently pursuing the B.S. degree in computer science with Guangdong University of Technology (GDUT), Jieyang, China. His research interests include machine vision and intelligent inspection.

YUE CHU is currently pursuing the degree in computer science (machine vision and intelligent inspection) with Guangdong University of Technology, Jieyang, China. Her academic pursuits are deeply rooted in the innovative application of technology to enhance visual recognition systems and improve automated inspection processes.

ZHEN LE was born in Guangzhou, China, in 2002. He is currently pursuing the degree in electronic information engineering with Guangdong University of Technology. His research interests include embedded systems and signal processing.

YELING XU is currently pursuing the degree in artificial intelligence with Guangdong University of Technology. His research interests include computer vision and multimodal large language models. He is dedicated to researching explainable artificial intelligence and promoting its applications across various industries.

JIANFENG LI has a wealth of experience in designing and developing advanced communication systems in leading enterprises. He is currently a dedicated Faculty Member with Guangdong University of Technology. His research interests include broadband communication systems, signaling, machine vision, and artificial intelligence.