An Improved Multi-target Detection Algorithm in UAV Aerial Images Based on YOLOv8s Framework

Changyou Wang, Qing Zhang, and Jie Huang

***Abstract*—In UAV aerial photography scenarios, target detection faces numerous challenges, particularly the issues of detecting overly small target sizes and high background similarity of target images, which severely limit detection accuracy and efficiency. To address these challenges, this paper proposes a depth-optimized target detection algorithm based on the YOLOv8s framework. The core innovation of this algorithm lies in the multi-dimensional enhancement of the YOLOv8s model, aiming to substantially improve detection performance for small targets and in complex background environments. First, we introduce a convolutional module that incorporates sensory wild attention, replacing the traditional convolutional layer in the backbone network. This design effectively enhances the model's learning efficiency for detailed features by boosting the network's capacity to capture local region and global context information, thereby accelerating the speed and accuracy of target detection. Secondly, we innovatively incorporate a multi-faceted attention mechanism within the subsequent feature extraction phase of the backbone network, which is subsequently followed by a pooling layer. This refinement significantly enhances the precision of our algorithm compared to its original counterpart. This mechanism can dynamically adjust feature weights to strengthen the representation of key target features while suppressing background noise, which significantly reduces the false detection rate, especially for small targets and complex backgrounds. Additionally, we have made pivotal improvements to the neck network structure by replacing the original C2f module with a more efficient ELAN (Enhanced Local Aggregation Network) module. The ELAN module learns directly from the original feature map, minimizing information loss and enhancing feature representation, thereby further impr- oving detection accuracy.**

***Index Terms*—YOLOv8s, Multi-target detection, Feel the wild attention, UAV aerial images, Loss function optimization**

I. INTRODUCTION

UAVs (Unmanned Aerial Vehicles) have a diverse range of applications in fields such as agricultural production, pedestrian detection, and military combat [1], with target detection in aerial images occupying a pivotal position in the utilization of UAVs. Due to the UAV's high relative altitude above ground and the extensive field of view of aerial imagery, issues arise such as lower pixel availability for target objects, smaller sizes, and increased recognition difficulty [2]. Target detection algorithms are classified into traditional detection algorithms and convolutional neural network (CNN) algorithms based on deep learning. As traditional target detection algorithms exhibit poor anti-interference capabilities and are unsuitable for complex detection environments, CNN algorithms based on deep learning are more frequently employed in UAV aerial photography. These algorithms offer more accurate and faster detection and are currently the mainstream approach. Deep learning-based target detection algorithms are primarily categorized into two-stage and single-stage algorithms. Two-stage target detection algorithms, exemplified by R-CNN [3], require auxiliary sub-networks to generate candidate bounding boxes, resulting in slower detection speeds. In contrast, single-stage target detection algorithms, which include single-shot detectors (SSD [4]) and the YOLO [5] series, directly classify and regress feature maps, thereby offering a broader range of application scenarios. Zhang [6] proposed replacing the CIoU loss function with a more efficient EIoU loss function to directly minimize the difference between the width and height of the target and anchor boxes, achieving faster convergence and improved localization accuracy. He [7] integrated channel and spatial attention by embedding a lightweight feature-enhanced backbone network with Shuffle Channel and Spatial Attention modules within the backbone of YOLOv3. This enhancement improved accuracy, reduced model complexity, and accelerated small object detection. Hang [8] introduced an additional small target detection head in YOLOv5 to retain as much feature information as possible for small-sized targets, replacing the original convolutional prediction heads with Swin Transformer Prediction Heads (SPHs). Ranjai [9] incorporated a probe head and ConvMixer to extract one-to-one feature relationships using deep and pointwise convolutions, which facilitates better handling of small target object tracking by establishing spatial and channel relationships. Liu [10] simplified the model and enhanced object detection efficiency by introducing a DCS layer to replace the original convolutional block. Consequently, they proposed a lightweight Slim-BiFPN to replace the original Feature Pyramid Network (FPN) in YOLOv5. Yang [11] integrated the GhostNet module, modified the loss function, and redesigned anchor boxes based on YOLOv5 to improve small target detection accuracy. To further

Manuscript received October 2, 2024; revised February 14, 2025.

This work is supported by the General Projects of Local Science and Technology Development Funds Guided by the Central Government (Grant No: 2023ZYD0001) and the Natural Science Foundation of Sichuan Province (Grant no. 2023NSFSC007).

Changyou Wang is a professor of College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan 610225 China. (Corresponding author, e-mail: wangchangyou417@163.com).

Qing Zhang is a postgraduate student of College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan 610225 China. (e-mail: 470053304@qq.com).

Jie Huang is a postgraduate student of College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan 610225 China. (e-mail:1594278623@qq.com).

refine the detection of various object types, YOLOv6, YOLOv7, and YOLOv8 algorithms have been continuously developed on the foundation of YOLOv5, as detailed in references [12-16].

Although YOLOv8, as one of the current state-of-the-art target detection algorithms, has demonstrated excellent detection performance, it still faces significant challenges when dealing with detection tasks involving small targets, complex backgrounds, and variable lighting conditions. In this paper, these challenges are deeply analyzed, and a series of innovative improvement strategies are proposed. By embedding a reference-free attention module in the backbone network to enhance the comprehensiveness of feature extraction, fusing sense-field attention into the convolutional layer to improve multi-scale target recognition capability, replacing the C2f module of the necking network with an Elan module to enhance the model's robustness, and optimizing the training process by using a WIoU loss function to eliminate the interference of low-quality samples, this paper successfully constructs a more adaptable target detection model with better performance. Experimental results show that the improved algorithm achieves significant performance improvements in several evaluation dimensions.

II. The Algorithmic Scheme of This Paper

A. *Original YOLOv8 Model*

The YOLO series, introduced in 2016, has been renowned for its rapid response speed and high accuracy. In January 2023, the YOLOv5 team at Ultralytics proposed the YOLOv8 [17] model, which incorporates several improvements over YOLOv5 and offers versions such as s, m, n, l, and x. YOLOv8 consists of three parts: the backbone network (Backbone), the neck network (Neck), and the detection head (Head). Its structural model is illustrated in Figure 1. Compared to YOLOv5, YOLOv8 features notable modifications, including: firstly, replacing the C3 structure in the backbone network of YOLOv5 with the C2f structure; secondly, adopting the SPPF structure to further enhance feature semantic expression ability while preserving original features and improving the information richness of the pooling layer; the improvements in the neck network are similar to those in the backbone network, with the convolution kernel of the first convolutional layer being optimized, and C3f being replaced with the same C2f structure as in the backbone network, while redundant connection layers are removed and the number of different channels is adjusted according to the image size; additionally, the coupled head of YOLOv5 is changed to a decoupled head, and the anchor-based detector is transformed into an anchor-free detector; finally, Distribution Focal Loss fused with CIoU Loss is introduced to calculate the bounding box loss.

B. *Improved YOLOv8 Model*

In the context of target detection models for Unmanned Aerial Vehicles (UAVs), the majority of aerial images, due to the shooting angles and high altitudes involved, encompass small and densely packed targets. These targets, coupled with the intricate and variable backgrounds filled with interfering factors, render the YOLOv8 algorithmic model susceptible to misdetections, omissions, and other related issues in this specific application scenario.

In this paper, a series of enhancements are introduced to the foundational YOLOv8s model to effectively tackle the challenge of low detection accuracy in aerial imagery using this algorithm. Firstly, multiple attention mechanisms are integrated into the backbone network [18], enabling the model to concentrate on the target region, mitigate background interference, and seamlessly integrate spatial, channel, and temporal attention to bolster its adaptability. Secondly, the Elan module [19] is utilized to replace the C2f feature extraction module, thereby facilitating more comprehensive and granular feature fusion by augmenting channel dimensions and increasing the number of branches to capture feature information across various levels and scales. Following this, within the backbone network, the standard convolution is substituted with the Fusion Sensory Wild Attention (FSWA) convolution module [20], which enhances the capacity to focus on and extract pivotal feature regions by adaptively assigning weights. Lastly, given the high computational complexity of CIoU [21] Loss and DFL Loss, which struggle with complex and demanding scenes, Wise-IoU [22] is adopted as the loss function to elevate detection accuracy and demonstrate superior adaptability to a diverse range of scenarios. The resultant algorithm model architecture after these modifications is depicted in Figure 2.

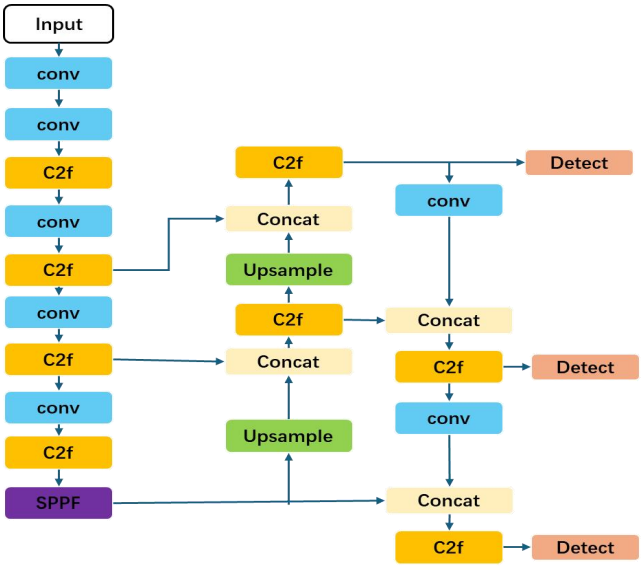

Fig. 1. YOLOv8 schematic diagram

C. *Add Triplet Attention Module*

In recent years, the attention model has been increasingly utilized in the field of deep learning algorithms. The attention model mimics the human visual attention mechanism, allowing the algorithmic model to focus more on the object features in target detection tasks, thereby facilitating the realization of these tasks. Triplet Attention is a mechanism designed for processing ternary (three-dimensional) channel data attention. It accelerates the model's ability to process data and quickly locate the target object by establishing relationships among the three dimensions: space, height, and width. Triplet Attention not only establishes weight relationships between height and channel, width and channel, but also determines the final attention weight relationship among all three dimensions through operations such as rotation, rearrangement, and convolution. This enables the model to pay greater attention to multi-dimensional features. The principle of Triplet Attention in the context of three-dimensional channels is illustrated in Figure 3.

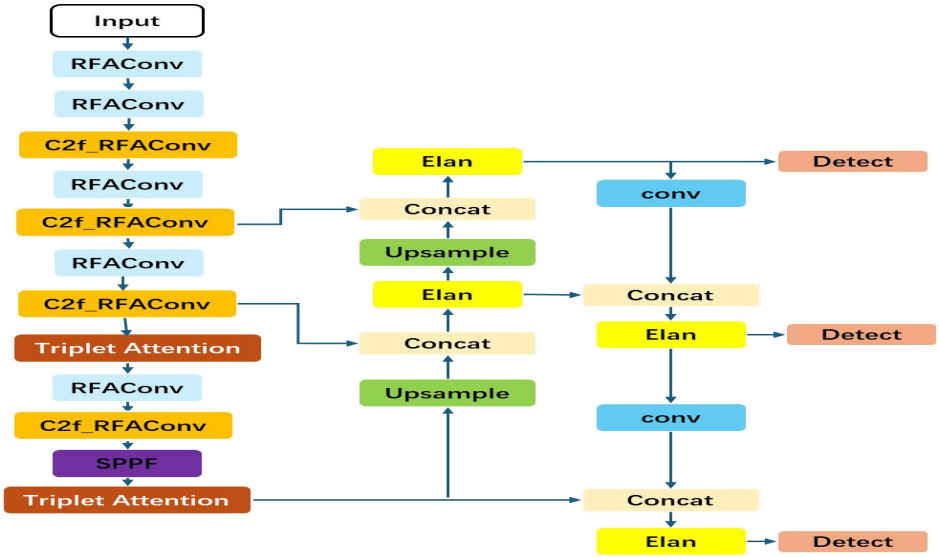

Fig. 2. Structure of the improved YOLOv8 algorithm

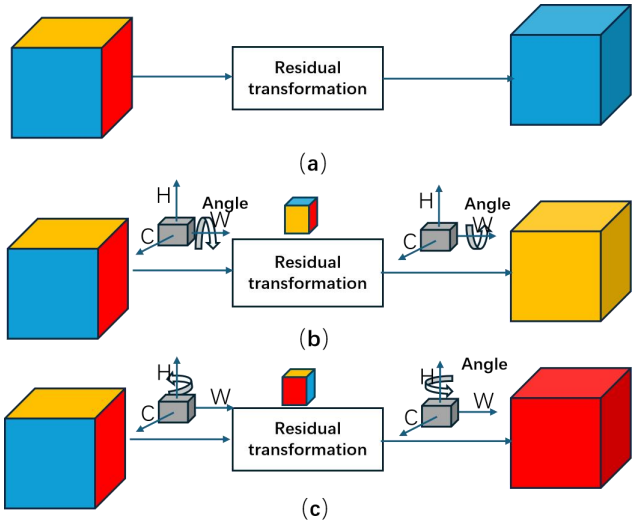

Fig. 3. Schematic diagram of the multiple attention mechanism

Triplet Attention constructs a ternary attention mechanism with three branches. One branch performs direct z-pooling compression on the original tensor to filter key channel features and reduce complexity. The weights are generated through convolution, batch normalization, and a Sigmoid activation function, which emphasize the important original features and act upon the original tensor output. The b-branch rotates the tensor 90 degrees counterclockwise along the W-axis to form a rotated tensor, facilitating the discovery of connections between horizontal features. The rotated tensor undergoes z-pooling compression and convolution with batch normalization to obtain and optimize important horizontal features and data. Sigmoid generates weights to focus on key horizontal features, which are then applied to the rotated tensor output to enhance horizontal orientation processing.

For branch c, the channel of the original input tensor is compressed by Z-pooling to a tensor of shape (2, H, W), and the input tensor is rotated along the height axis, transforming the viewpoint to capture vertical feature relationships. The rotated tensor is then compressed and downscaled by Z-pooling, and key vertical features are extracted. These features are subsequently optimized for feature and data distribution through convolution and batch normalization. Attention weights are generated by the Sigmoid function, which are applied to the rotated tensor output to reinforce the vertical direction of attention. The tensors of shape (2, H, W) generated by each branch are aggregated by simple averaging. The algorithmic model presented in this paper chooses to add two Triplet Attention modules to the backbone network.

D. Improved Feature Extraction Module

The neck network of YOLOv8 utilizes CSPDarknet to extract features from the input image. The C2f module in YOLOv8 comprises CSP and FFM, which constitutes the core component of the feature extraction process for the YOLOv8 neck network. The specific structure of this module is shown in Figure 4.

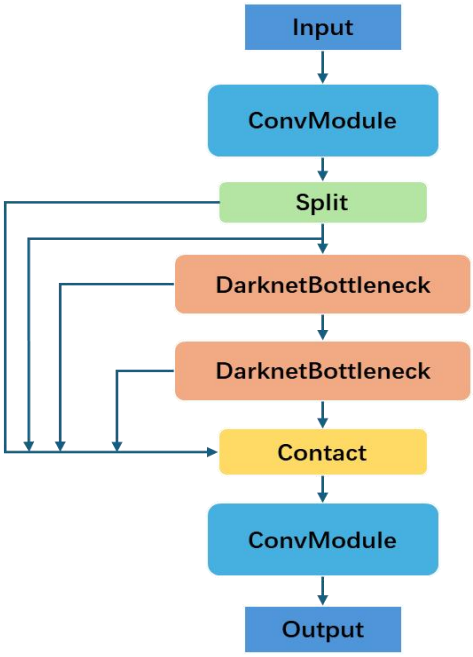

Fig. 4. C2f schematic

Although C2f in YOLOv8 employs a simpler convolution method to streamline the model, the resultant loss of gradient information still leads to poor tracking performance. To improve the tracking accuracy of YOLOv8, a module with reduced gradient information loss is selected to replace the C2f module. The ELAN module in YOLOv7 directs different groups of computational blocks to learn

more diversified features by regulating the shortest and longest gradient paths, enabling the deep learning network to learn and converge more efficiently. The structure of this module is shown in Figure 4.

ELAN adds several sets of convolutional modules to the original gradient transmission path, augmenting the foundation of the original image features. It reorganizes and merges features from different layers to enhance the image features. There are a total of four paths for feature extraction in ELAN. The first path involves passing through a 1x1 convolution module for dimensionality reduction. The second path achieves feature connection and fusion between cross-layer strata, first passing through a 1x1 convolution layer for dimensionality reduction and then through four 3x3 convolution modules for feature extraction. The third path obtains feature results through two 3x3 convolution modules. The fourth path involves passing through four 3x3 convolution modules for dimensionality reduction and feature processing. Finally, the four sets of features are summed to extract the final result.

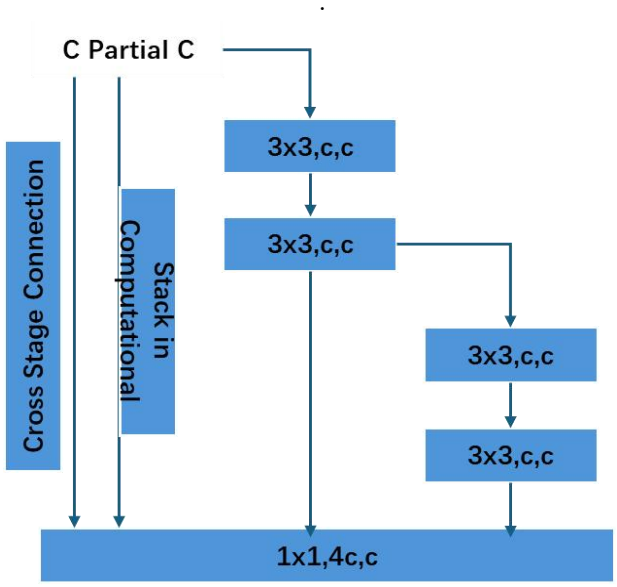

Fig. 5. ELAN schematic diagram

Replacing the C2f module in YOLOv8 with the ELAN module as the feature extraction backbone network, without incurring gradient loss, increases the accuracy of the network.

E. Improved Feature Extraction Module

In the backbone network of YOLOv8, the regular convolution module and C2f are primarily used for extracting network features. To better enable the network to focus on important features and enhance the model's anti-interference ability, the Receptive Field Attention Convolution (RFAConv) is selected to replace the regular convolution module. Receptive field attention convolution involves replacing the regular convolution operation with the RFA convolution operation, which not only attends to the spatial features of the receptive field but also provides effective attention weights for large-sized convolution kernels. Its schematic diagram is shown in Figure 6.

When the input feature vector $X \in R^{C\times H\times W}$ is changed to a sensory wild space feature of dimension $9C \times H \times W$ after Group conv, where C, H, and W represent the number of channels, height, and width of the input pixel. When the image undergoes average pooling to aggregate the global information of each receptive field feature, then the group convolution operation is used to interact, and finally softmax is used to emphasize the significance in each receptive field feature, the formula F for the receptive field is follow

$$F = Soft\max(g^{1\times1}(AvgPool(X)))\times \\ \mathrm{ReLU}(Norm(g^{k\times k}(X))) = A_{rf}\times F_{rf}, \quad (1)$$

where $g^{1\times1}$ represents a grouped convolution of size 1×1, k represents the size of the convolution kernel, Norm stands for normalization, X stands for the input feature map, Norm represents the normalization, and ReLU is the activation function. A_{rf} refers to the attention diagram, and F_{rf} is the spatial feature of the changed receptor field. F is obtained by multiplying A_{rf} and F_{rf}, which helps the network to better process the information in the receptive field space according to the importance weight of features, and improves the adaptability of the network.

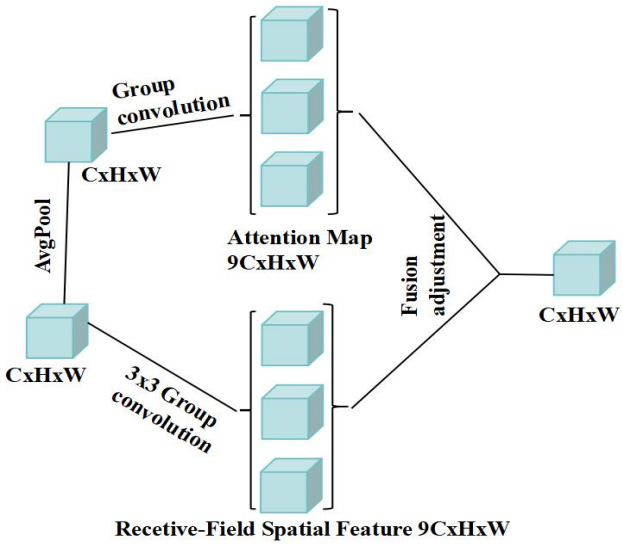

Fig. 6. Structure of RFAConv

F. Loss Function Improvement

The YOlOv8s model uses CIoU Losses and DFL Losses to predict the bounding box regression losses, CIoU Losses can quickly reduce the distance between two boxes and can converge faster for cases where the target is duplicated or the target is incomplete resulting in detection of objects with extreme aspect ratios. However, the convergence speed of CIoU decreases sharply when the target object has diverse shapes or the target has a large scale change due to fast speed movement. For this reason, Wise-IoU [20] is adopted as the loss function in this paper.

Wise-IoU constructs a two-layer attention mechanism in order to solve the problem of increasing the training intervention and decreasing the penalty when the anchor box and the target box overlap, and the calculation formulas are as follows:

$$L_{WIoUv1} = R_{WIoU}\times L_{IoU}, \quad (2)$$

$$L_{IoU} = 1 - IOU, \quad (3)$$

and

$$R_{WIoU} = \exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(w_g^{\ 2}+H_g^{\ 2})}\right), \quad (4)$$

where IOU denotes the intersection ratio of the prediction box and the real box, L_{IoU} is the boundary frame loss, and R_{WIoU} is the constructed distance attention. x and y are the horizontal and vertical coordinates of the center point of the prediction box; x_{gt} and y_{gt} are the horizontal and vertical coordinates of the center point of the real box; W_g and H_g are the width and height of the minimum external rectangle of the prediction box and the real box. Then, according to the loss function L_{WIoUv1} of the constructed two-layer attention mechanism, a nonmonotonic focusing coefficient is constructed to obtain the Wise-IoU v2 boundary frame loss. The calculation formula is as follow:

$$L_{WIoUv2} = r \times L_{WIoUv1}, \tag{5}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\alpha}}, \tag{6}$$

and

$$\beta = \frac{L^*_{IoU}}{\overline{L_{IoU}}} \in [0, +\infty), \tag{7}$$

where β is the outlier degree, representing the anomaly degree of the prediction box, $\overline{L_{IoU}}$ is the average value of L_{IoU}, and L^*_{IoU} is the monotonic focus coefficient which effectively reduces the weight of simple examples in the loss value. δ and α are hyper parameters, and the gradient gain r can be obtained by formula (6), which is applied to L_{WIoUv1} to obtain the Wise-IoU v2 bounding frame loss.

III. Experimental Setting and Analysis of Results

A. Experimental Environment

The dataset utilized in this paper is the DOTAv1.5 dataset. The DOTAv1.5 dataset has been meticulously cleaned and processed, and it is divided into 14,384 images for the training set and 4,874 images for the test set. The DOTA dataset encompasses 16 categories, including airplanes, ships, storage tanks, baseball stadiums, tennis courts, basketball courts, ground-level runways, harbors, bridges, small vehicles, large vehicles, helicopters, roundabouts, soccer fields, swimming pools, and container cranes. The experimental environment for this paper is as follows: the host system is Ubuntu 21.04, the computer GPU is an NVIDIA RTX 3090, the RAM is 16GB, the development language is Python 3.16, and the framework used is PyTorch 2.0.0.

B. Data sets and assessment indicators

This experiment employs Mean Average Precision (mAP) to assess the detection performance of the algorithm. mAP encompasses two metrics: mAP@0.50 and mAP@0.50:0.95, which quantify the degree of overlap between the target detection box and the ground truth box. Specifically, mAP@0.50 indicates the average precision of all classes when the Intersection over Union (IOU) threshold is set to 0.5, whereas mAP@0.50:0.95 reflects the average precision across a range of IOU thresholds from 0.5 to 0.95. A detection is considered a True Positive (TP) if its IOU with the ground truth frame exceeds the specified IOU threshold; otherwise, it is classified as a False Positive (FP). False Negatives (FN) are determined by subtracting the number of TPs from the total number of ground truth frames. Precision is the ratio of TPs to the total number of samples predicted as positive by the model, reflecting the accuracy of positive predictions. Recall, on the other hand, is the ratio of TPs to the actual number of positive samples present. The calculation formulas for mAP, Precision, and Recall are provided as follows:

$$\text{mAP} = \frac{1}{N} \sum_{n \in N} AP(n), \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{9}$$

and

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{10}$$

In this paper, mAP and Precision and Reca are selected as the evaluation models to evaluate the effectiveness of the model in UAV application scenarios.

C. Comparison with Baseline Model

To validate the improvements of the model introduced in this paper compared to the benchmark model, we trained both the enhanced YOLOv8s model and the original YOLOv8s model independently for 280 epochs using identical parameters, and subsequently evaluated them on the validation set. Figure 7 depicts the variations of certain key metrics with the number of training epochs during the training process of both the enhanced YOLOv8s model and the original YOLOv8s model. As observed, the yellow line represents the data from the model presented in this paper, while the blue line corresponds to the data from the original YOLOv8s model. Upon completion of the experiment, the yellow line data progressively surpasses the blue line data as the number of epochs increases, indicating that the model proposed in this paper outperforms the benchmark model in terms of average accuracy and detection rate.

D. Ablation Experiment

To verify the effectiveness of the enhanced modules, YOLOv8 is employed as a benchmark model, with improvements made to its backbone network, feature extraction module, loss function, and various other components. The first through fifth rows in Table I outline the specific operations: optimizing the loss function within the model, replacing the feature extraction network of the neck network, introducing the Triplet Attention module, and integrating the Convolutional Sensory Field Attention (C2f_RFAConv), respectively. Based on the experimental data presented in the table, the following conclusions can

be drawn: Firstly, after optimizing the YOLOv8s loss function to Wise-IOU, the model's mAP@0.50 increased by 0.7%, and its mAP@0.50:0.95 increased by 0.3%. Secondly, incorporating the ELAN module into the neck network resulted in a 0.1% increase in the model's mAP@0.50 and a 0.4% increase in its mAP@0.50:0.95. Subsequently, introducing two Triplet Attention modules into the backbone network led to a 1.2% increase in the model's mAP@0.50 and a 0.4% increase in its mAP@ 0.50:0.95. Lastly, integrating the C2f_RFAConv module into the backbone network boosted the model's mAP@0.50 by 0.2% and its mAP@0.50:0.95 by 0.6%. Compared to the baseline YOLOv8 algorithm model, the improvements proposed in this paper result in a 1.7% increase in both mAP@0.50 and mAP@0.50:0.95. Although there is a 10.6% increase in GFLOPs, the accuracy exhibits a significant improvement, and the model's real-time processing capability remains unaffected.

E. Comparison experiment

To validate the efficacy of the enhanced YOLOv8 model, this paper conducts comparative experiments with other prevalent models, all trained on the DOTA V1.5 dataset to assess their performance metrics. The models included in the comparison encompass Faster-RCNN [23], SSD [4], RetinaNet [24], YOLOv3 [25], YOLOv5s and YOLOv5m [26], YOLOv8s, as well as the improved YOLOv8 model presented in this paper, as outlined in Table II. The comp -arative experiment results presented in the table unequivocally demonstrate that, in contrast to other models, the model proposed in this paper exhibits improvements in Precision, Recall, and mAP@0.50. Notably, there are significant enhancements in both the recall rate and precision rate. Given the high altitude of aerial photography, the majority of detection targets in this dataset are small. Faster-RCNN and SSD, being two-stage detection algorithms, suffer from inadequate feature extraction for small targets, coupled with their intricate overall structure and high computational demands, leading to suboptimal detection performance. While YOLOv3 represents an improvement over two-stage detection methods, it involves numerous downsampling operations, which tend to result in the loss of features of small targets. Furthermore, due to its limited scale level, YOLOv3 lacks the fineness and comprehensiveness required for feature extraction and fusion, resulting in less than optimal detection performance. Finally, when comparing the models YOLOv5s, YOLOv5m, YOLOv8s, and other mainstream one-stage models in this paper, the model proposed in this paper attains the highest levels of accuracy, detection precision, and recall, showcasing exceptional overall performance.

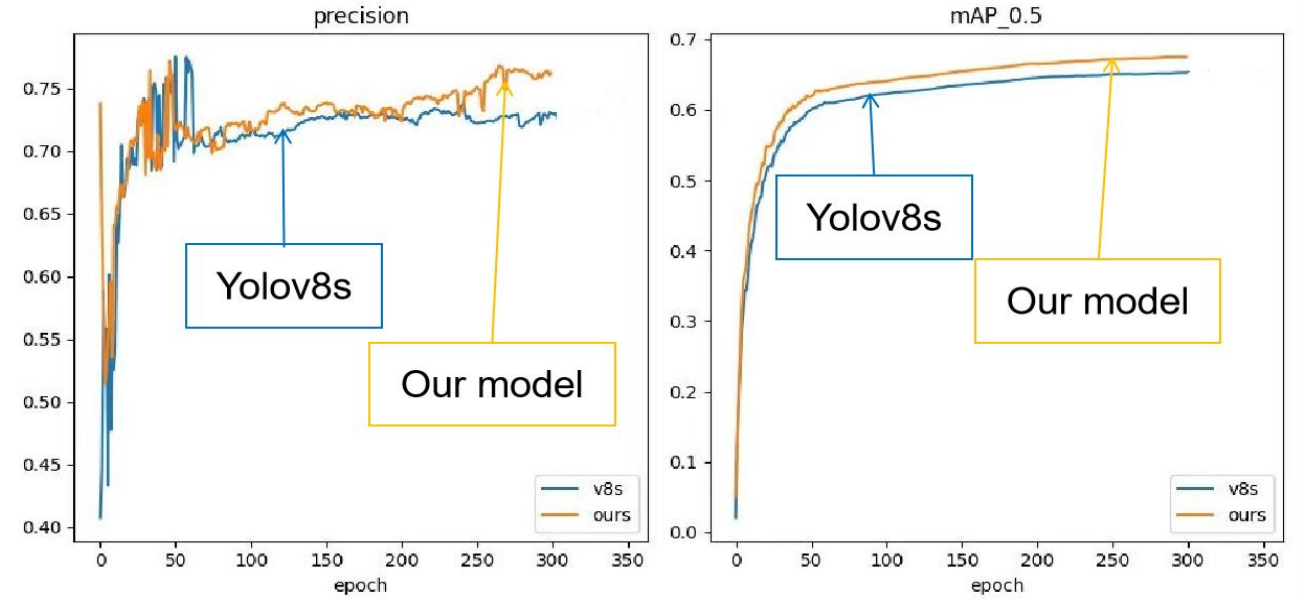

Fig. 7. Comparison chart of experimental results

TABLE I
ABLATION EXPERIMENT DATA

Comparison of each parameter of ablation experiment target detection				
Models	mAP@0.50	mAP@0.50:0.95	Parameters	GFLOPs
YOLOv8s	65.7	43.7	11.13M	28.5
YOLOv8s-WIOU	66.4	44	11.13M	28.5
YOLOv8s-WIOU-ELAN	66.5	44.4	14.79M	34.9
YOLOv8s-WIOU-ELAN-Triplet Attention	67.2	44.8	15.66M	38.1
YOLOv8s-WIOU-ELAN-Triplet Attention-RFAConv	67.4	45.4	15.78M	39.1

TABLE II
COMPARATIVE EXPERIMENTAL DATA

Models	P	R	mAP@50
Faster-RCNN	67.5	56.4	59.9
SSD	66.1	49.6	53.3
RetinaNet	67.8	55.5	59.0
YOLOv3	67.9	51.8	55.7
YOLOv5s	73.3	60.4	65.8
YOLOv5m	73.4	62.5	67.5
YOLOv8s	74.9	60.0	65.7
Ours	75.8	61.7	67.4

F. Ablation Experiment

The comparative analysis, grounded in real-world scenario applications, is presented hereinafter. As depicted in Figure 8, the sequence from left to right showcases the labeled information intended for detection in the original image, the detection outcome yielded by the original YOLOv8 algorithm model, and the detection result obtained using the model introduced in this paper. Upon scrutiny, it becomes apparent that while seven small vehicles were originally present for detection, the original model only identified five. In stark contrast, the model proposed herein successfully detected all seven small vehicles, markedly alleviating the issue of missed detections. Furthermore, it augmented the confidence score of the leftmost small vehicle from 0.4 to 0.6.

Figure 9(a) illustrates a multi-target category scenario, necessitating the identification of 7 small vehicles, 2 ports, and 4 ships. Upon examining Figure 9(b), it is evident that the original algorithm model detected 8 small vehicles (including one false positive), 2 ports, and 2 ships, missing 2 ships due to their minuscule size. Conversely, our proposed model accurately identified all targets, as exhibited in Figure 9(c).

Figure 10(a) presents a detection scenario where both large and small targets coexist, requiring the identification of 3 airplanes and 1 small vehicle. In contrast, Figure 10(b) reveals that the original algorithm model predominantly focused on detecting large targets, neglecting the small target object. However, Figure 10(c) underscores that the model introduced in this paper precisely detected the overlooked small vehicle.

Figure 11 contrasts the experimental results, underscoring the enhanced capability in detecting scenes with multi-sized objects in complex environments. In Figure 11(a), the objectives encompassed identifying 1 swimming pool, 7 baseball diamonds, 1 tennis court, and 1 ground track field. The output from the original algorithm model in Figure 11(b) not only failed to detect one baseball diamond but also erroneously detected a soccer field instead of a swimming pool, notably missing the swimming pool in the upper left corner. In sharp contrast, the model proposed in this paper accurately detected all targets, as delineated in Figure 11(c).

Upon scrutinizing and comparing these four sets of images, it becomes unequivocally clear that the original YOLOv8s model exhibits instances of missed detections and false positives. In stark contrast, the refined model introduced in this paper demonstrates an unerring ability to correctly detect all targets. Additionally, in Figure 8, the confidence score for the small vehicle in image (c) surpasses that in image (b). Similarly, in Figures 9, 10, and 11, the confidence scores for the identified targets in image (c) are consistently higher than those in image (b). This comparison conclusively attests to the superior accuracy of the algorithm presented in this paper compared to the original YOLOv8s algorithm.

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

In this paper, we have presented a novel algorithmic model aimed at addressing the prevalent issues of leakage and misdetection in unmanned aerial vehicle (UAV) imagery. Leveraging the foundational strengths of the YOLOv8 model, our contributions lie in the introduction of several innovative components. Specifically, we have integrated the triplet attention mechanism and the C2f_RFAConv module into the backbone network to enhance feature extraction capabilities. Furthermore, we have incorporated the ELAN module into the neck network to facilitate more effective feature fusion. Additionally, we have adopted the Wise-IoU loss function as a replacement for the original loss function to improve the precision of object detection.

Through extensive experiments conducted on the DOTAv1.5 dataset, our results demonstrate that the proposed improved algorithm achieves notable performance enhancements across multiple evaluation dimensions. When compared to other currently popular object detection algorithms tailored for aerial imagery captured by drones, our model exhibits superior performance, highlighting its potential for addressing the challenges associated with UAV imagery.

B. Innovation Points

(1) Integration of the Triplet Attention Mechanism and C2f_RFAConv Module: By introducing these components into the backbone network, we have significantly improved the model's ability to capture and represent complex features within UAV imagery. This enhancement is crucial for detecting small or occluded objects, which are common in aerial photography.

(2) Incorporation of the ELAN Module: The integration of the ELAN module in the neck network allows for more robust and efficient feature fusion. This improvement leads to better generalization and adaptation of the model to various scenes and lighting conditions, which are typical in UAV-captured imagery.

(3) Adoption of the Wise-IoU Loss Function: By replacing the original loss function with the Wise-IoU loss, we have achieved more precise bounding box predictions.

This refinement is essential for improving the overall accuracy and reliability of object detection in UAV imagery.

C. Application Value

The proposed algorithm has significant application value in various domains, including but not limited to military reconnaissance, disaster response, and urban planning. By providing a more accurate and reliable object detection capability, our model can enhance the operational efficiency and decision-making capabilities of UAV-based systems. For instance, in military reconnaissance, accurate detection of targets can provide critical information for strategic planning and mission execution. In disaster response, rapid and accurate identification of affected areas and resources can facilitate more effective emergency management and resource allocation.

D. Future Work

While the proposed algorithm demonstrates promising results, there is still considerable room for improvement in real-time performance and target detection accuracy. Future research will focus on optimizing the model's architecture and training process to further enhance its capabilities. Specifically, we plan to explore more advanced attention mechanisms and convolutional modules to improve feature extraction and representation. Additionally, we will investigate more efficient loss functions and optimization strategies to refine the bounding box predictions and reduce computational overhead.

Moreover, we will expand the experimental dataset to include more diverse and challenging scenarios to evaluate the robustness and generalization of the proposed model. This will involve collecting and annotating UAV imagery from various sources and environments, ensuring that the model can perform consistently well across different contexts.

In conclusion, our work presents a significant step forward in addressing the challenges of UAV imagery- based object detection. By introducing innovative components and demonstrating promising results, we hope to inspire further research and development in this important area.

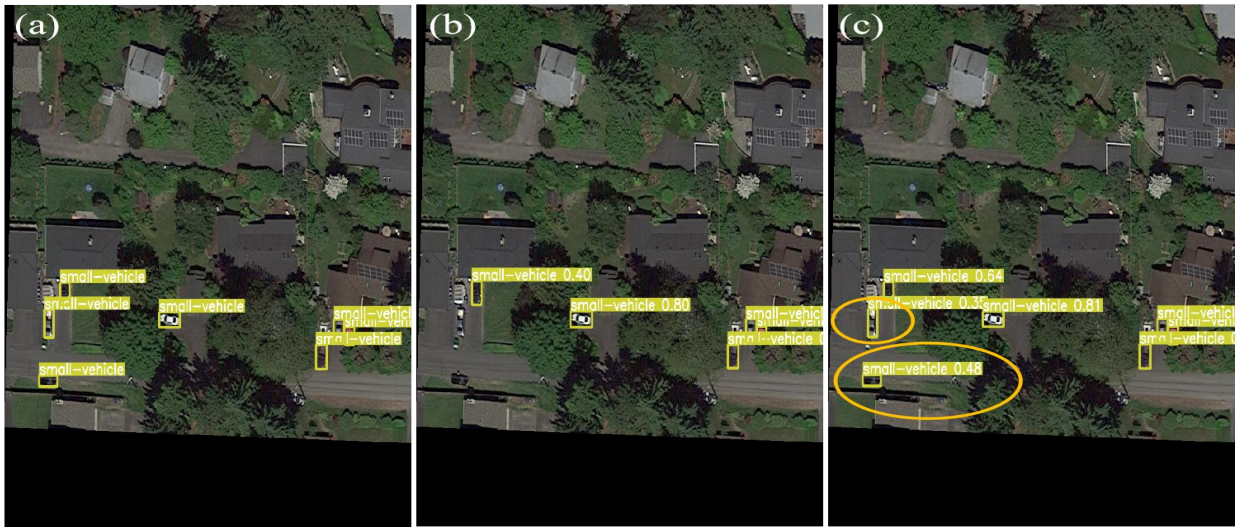

Fig. 8. Comparison of experimental results on the improvement of single-class target detection scenarios, where Figure (a) shows the target to be found in the original image, Figure (b) displays the detection results of YOLOv8s, and Figure (c) presents the detection results of the model proposed in this paper. The circle content is the main difference between our algorithm and YOLOv8s.

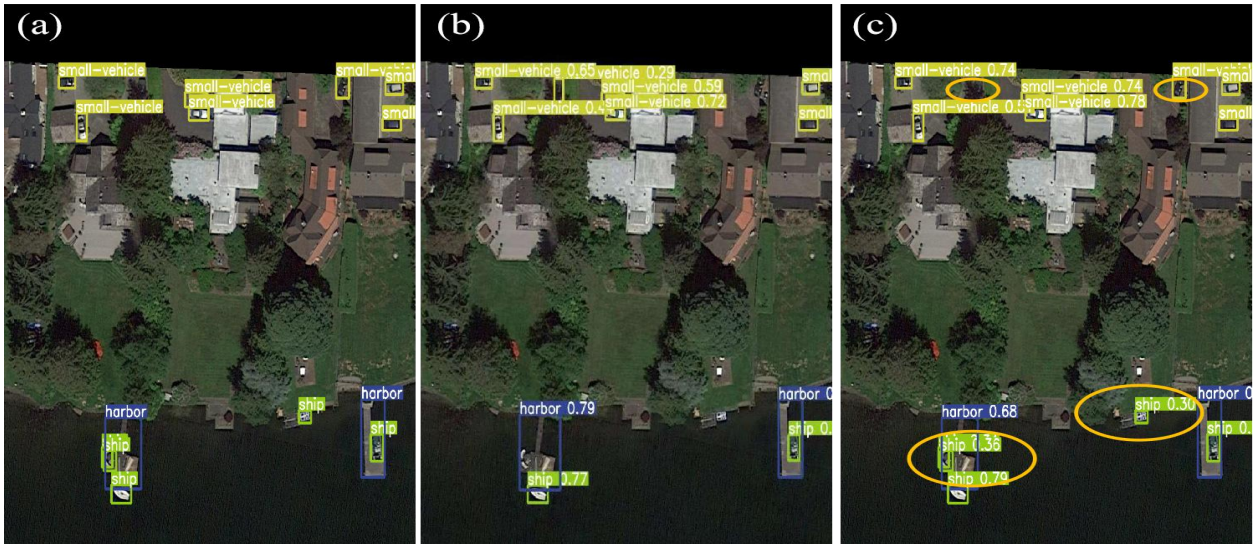

Fig. 9. Comparison of experimental results on the improvement of multi-class target detection scenarios, where Figure (a) shows the target to be found in the original image, Figure (b) displays the detection results of YOLOv8s, and Figure (c) presents the detection results of the model proposed in this paper. The circle content is the main difference between our algorithm and YOLOv8s.

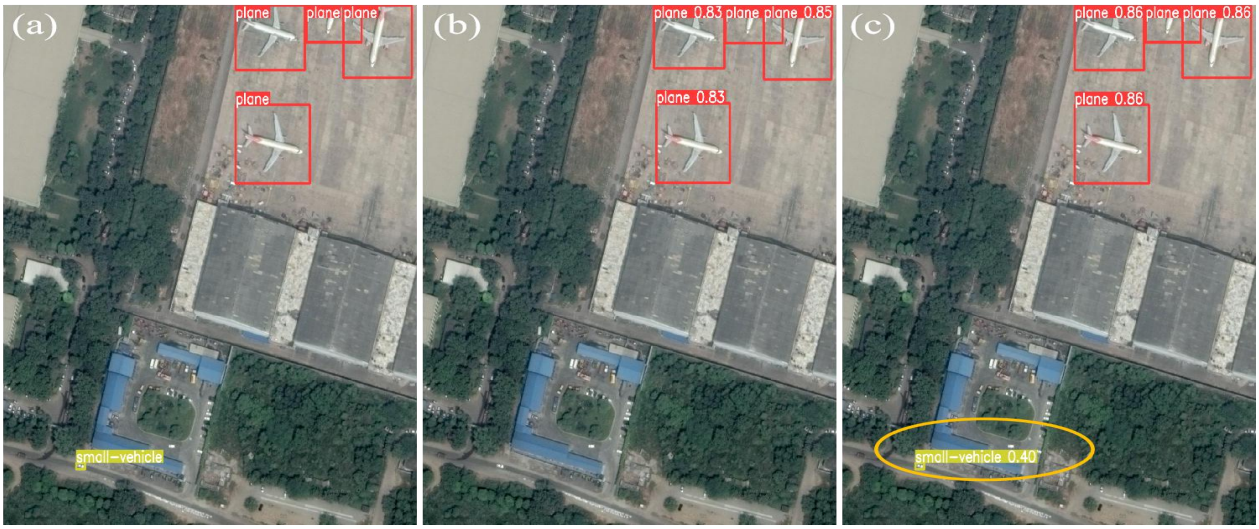

Fig. 10. Comparison of experimental results on the improvement of detection scenarios with coexisting large and small targets, where Figure (a) shows the target to be found in the original image, Figure (b) displays the detection results of YOLOv8s, and Figure (c) presents the detection results of the model proposed in this paper. The circle content is the main difference between our algorithm and YOLOv8s.

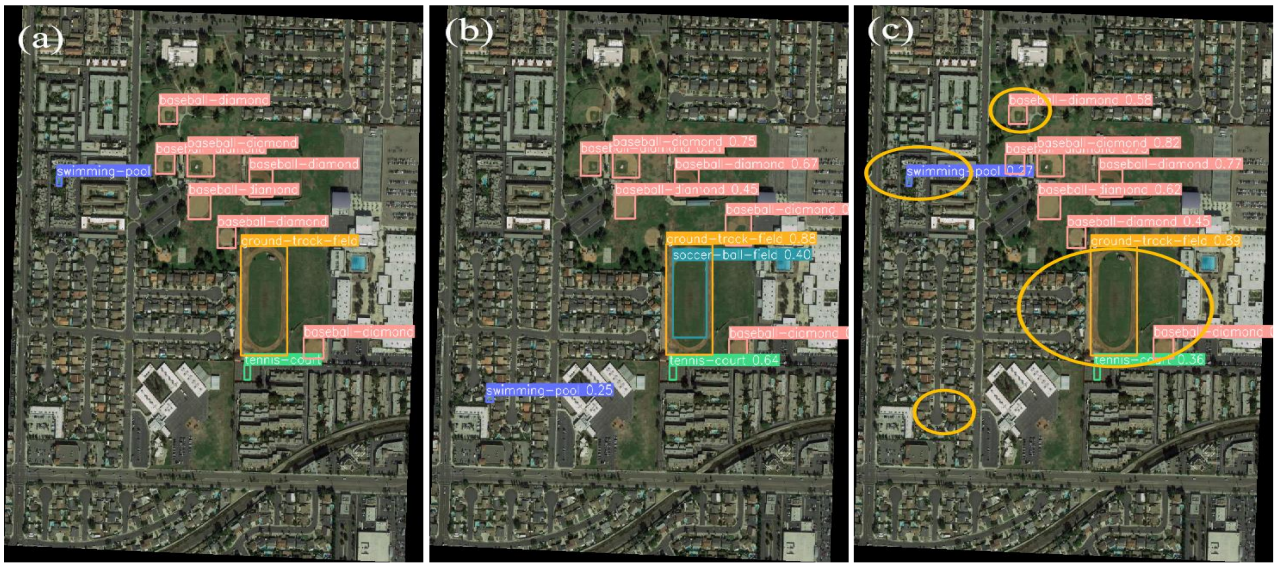

Fig. 11. Comparison of experimental results of the improvement effect of the detection scene with the coexistence of multi-size objects in a complex scene, where Figure (a) shows the target to be found in the original image, Figure (b) displays the detection results of YOLOv8s, and Figure (c) presents the detection results of the model proposed in this paper. The circle content is the main difference between our algorithm and YOLOv8s.

AUTHORS' CONTRIBUTIONS

C. Wang, Q. Zhang, and J. Huang contributed equally to each part of this work.

REFERENCES

[1] A. Bouguettaya, H. Zarzour, A. Kechida, et al., "A survey on deep learning-based identification of plant and crop diseases from UAV-based aerial images," Cluster Computing, vol.26, no.2, pp. 1297-1317, 2023.

[2] Q. Zhang, H. Zhang, X. Lu, "Adaptive feature fusion for small object detection," Applied Sciences, vol.12, no.22, Article ID: 11854, 2022.

[3] A. Y. Virasova, D. I. Klimov, O. E. Khromov, et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," Radioengineering, vol.85, no.9, pp.115-126, 2021.

[4] W. Liu, D. Anguelov, D. Erhan, et al., "Ssd: Single shot multibox detector," Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, pp. 21-37, 2016.

[5] J. Redmon, S. Divvala, R. Girshick, et al., "You only look once: Unified, real-time object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.779-788, 2016.

[6] Y. F. Zhang, W. Ren, Z. Zhang, et al., "Focal and efficient IOU loss for accurate bounding box regression," Neurocomputing, vol.506, pp. 146-157, 2022.

[7] X. He, R. Cheng, Z. Zheng, et al., "Small object detection in traffic scenes based on YOLO-MXANet," Sensors, vol.21, Article ID:7422, 2021.

[8] H. Gong, T. Mu, Q. Li, et al., "Swin-transformer-enabled YOLOv5 with attention mechanism for small object detection on satellite images," Remote Sensing, vol.14, no.12, Article ID:2861, 2022.

[9] R. Baidya, H. Jeong, "Yolov5 with convmixer prediction heads for precise object detection in drone imagery," Sensors, vol.22, Article ID:8424, 2022.

[10] C. Liu, D. Yang, L, Tang, et al., "A lightweight object detector based on spatial-coordinate self-attention for UAV aerial images," Remote Sensing, vol.15, no.1, Article ID:83, 2022.

[11] R. Yang, J. Zhang, X. Shang, et al., "Lightweight small target detection algorithm with multi-feature fusion," Electronics, vol.12, Article ID:2739, 2023.

[12] Y. W. Li, X. X. Zhang, "Object detection for UAV images based on improved YOLOv6," IAENG International Journal of Computer Science, Vol.50, no.2, pp.459-768, 2023.

[13] J. S. Fu, Y. Tian, "Improved YOLOv7 underwater object detection based on attention mechanism," Engineering Letters, Vol.32, no.7, pp. 1377-1384, 2024.

[14] Z. F. Hu, F. Y. Li, J. X. Shen, "A semantic SLAM integrated with enhanced YOLOv7 target detection algorithm," Engineering Lett-

ers, Vol.32, no.10, pp.1909-1920, 2024.
[15] W. Z. Teng, H. G. Zhang, Y. J. Zhang, "X-ray security inspection prohibited items detection model based on improved YOLOv7-tiny," IAENG International Journal of Applied Mathematics, Vol.54, no.7, pp.1279-1287, 2024.
[16] Y. F. Chai, X. X. Zhang, "Enhanced chest CT detection of pulmonary nodules based on YOLOv8," Engineering Letters, Vol.32, no.12, pp. 2221-2231, 2024.
[17] JOCHER G. Ultralytics YOLOv8[EB/OL]. [2023.01.10.]. https://github.com/ultralytics/ultralytics,2023.01.10.
[18] D. Misra, T. Nalamada, A. U. Arasanipalai, et al., "Rotate to attend: Convolutional triplet attention module," Proceedings of the IEEE/ CVF winter conference on applications of computer vision, pp.3139-3148, 2021.
[19] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp,7464-7475, 2023.
[20] X. Zhang, C. Liu, D. Yang, et al., "Rfaconv: Innovating spatital attention and standard convolutional operation," arxiv preprint arxiv: 2304. 03198, 2023.
[21] Z. Zheng, P. Wang, D. Ren, et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," IEEE transactions on cybernetics, vol.52, no.8, pp. 8574-8586, 2021.
[22] Z. Tong, Y. Chen, Z. Xu, et al., "Wise-IoU: bounding box regression loss with dynamic focusing mechanism," arxiv preprint arxiv: 2301. 10051, 2023.
[23] S. Ren, K, He, R. Girshick, et al., "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol.2015, Article ID:28, 2015.
[24] T. Y. Lin, P. Goyal, R. Girshick, et al., "Focal loss for dense object detection," Proceedings of the IEEE international conference on computer vision, pp.2980-2988, 2017.
[25] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement," arxiv preprint arxiv:1804.02767, 2018.
[26] JOCHER G. YOLOv5[EB/OL]. [2022.09.05.]. https:// github.com/ ultralytics/yolov5

Changyou Wang is Professor of College of Applied Mathematics, Chengdu University of Information Technology, Sichuan, China, since September 2017, and is Professor and Director of Institute of Applied Mathematics, Chongqing University of Posts and Telecommunications, Chongqing, China, from November 2011 to August 2017. He acts as a reviewer for Mathematical Reviews for American Mathematical Society, since 2014. His research area include delay reaction-diffusion equation, functional differential equation, fractional-order differential equation, difference equation, biomathematics, control theory and control engineering, neural network, and digital image processing. His education background are as follows: (1) Ph.D. degree from College of Applied Science, Beijing University of Technology, Beijing, China, in September 2008-June 2012, majoring in applied mathematics; (2) M.S. degree from College of Mathematics and Software, Sichuan Normal University, Chengdu, China, in September 2001-June 2004, majoring in applied mathematics; (3) B.S. degree from College of Mathematics, East China University of Science and Technology, Nanchang, China, majoring in basic mathematics, in September 1987-June 1989.

Qing Zang is a postgraduate student of School of Applied Mathematics, Chengdu University of Information Technology, Sichuan, China. Her research interests include neural networks, image and video processing, and data analysis.

Jie Huang is a postgraduate student of School of Applied Mathematics, Chengdu University of Information Technology, Sichuan, China. His research interests include neural networks, image and video processing, and data analysis.