IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 July 2024, accepted 27 August 2024, date of publication 2 September 2024, date of current version 16 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3452716

RESEARCH ARTICLE

DMFF-YOLO: YOLOv8 Based on Dynamic Multiscale Feature Fusion for Object Detection on UAV Aerial Photography

XIAOYANG QIU, YAJUN CHEN, CHAOYUE SUN, JIANYING LI, AND MEIQI NIU
School of Electronic Information Engineering, China West Normal University, Nanchong 637009, China

Corresponding author: Yajun Chen (scnccyj@cwnu.edu.cn)

This work was supported by the China West Normal University Talent Fund under Grant 463177.

ABSTRACT With the rapid proliferation of drones across various domains, aerial target detection has become increasingly crucial. However, the targets in aerial images present challenges such as scale variation, small size, and density, leading to suboptimal performance of current detectors on aerial images. Based on the aforementioned challenges, we design an efficient aerial target detection algorithm called DMFF-YOLO. Specifically, to address the issues of small target size and scale variation, we design the DMFF neck structure, adding a small target detection head to tackle the small target size problem, using the DMC module to fuse different scale features for enriching detailed information, and employing the DSSFF module to construct a scale sequence space to solve the target scale variation problem. In the network backbone, we employ RFCBAMConv modules as downsampling layers, which interact with receptive-field features to mitigate the information disparity caused by positional changes and outperform traditional convolutional layers. Finally, we design the Soft-NMS-CIoU module to address the issue of suppressing adjacent boxes due to dense targets. On the VisDrone dataset, compared to the original algorithm, our method reduces the number of parameters by 31.1% while achieving an 11.7% improvement in mAP50. Extensive experiments on the VisDrone, DOTA, and UAVDT datasets demonstrate that the proposed algorithm performs well in aerial image detection tasks.

INDEX TERMS Multi-scale feature fusion, small object detection, UAV, YOLO.

I. INTRODUCTION

In the field of drones, target detection in aerial images is the foundation of various research. With the rapid development of artificial intelligence and mechanical manufacturing, drones, with their excellent imaging performance and flight capabilities, have aided drone technology [1], intelligent traffic monitoring [2], agriculture [3], and other fields, improving work efficiency and reducing manpower costs. Meanwhile, object detection and tracking serve as prerequisites for many subsequent visual tasks, making real-time and effective detection of objects in drone aerial images of significant research significance [4].

Compared to natural images, objects in drone images exhibit characteristics such as multiscale, small size, and dense distribution, posing significant challenges for aerial target detection [5]. This phenomenon is primarily attributed to four factors. Firstly, due to the varying flight altitudes of drones, images captured by drones undergo significant scale variations for the same target size. Secondly, these images contain small and large targets, with noticeable differences in scale among different targets, thereby increasing the difficulty of detecting all aerial targets. Moreover, the elevated position of the drone results in a lower pixel ratio of captured objects in images, leading to loss of object details and insufficient information, which hampers accurate object detection. Additionally, the elevated position of drones provides a wide field of view, resulting in the presence of numerous objects of different sizes arranged densely, thus making detection

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

prone to false positives and false negatives. These factors collectively impede the advancement of drone image detection technology. Mainstream object detection models are primarily trained on natural scenes and are based on the Backbone-Neck-Head architecture, including R-CNN series [6], [7], YOLO seriesy [8], [9], and DETR series [10], [11]. However, these models often perform poorly in aerial target detection. To address the challenges posed by drone imagery, many researchers have developed various detection models tailored to the characteristics of aerial images.

To address the multi-scale issues in aerial imagery, multi-scale feature fusion techniques utilize feature pyramid structures to obtain different feature maps for recognizing objects at various scales. Jocher et al. [12] introduced FPN-PAN structure in YOLOv5, which combines deep semantic features with shallow detail features by up-sampling high-level features to alleviate the impact of multi-scale targets. Kang et al. [13] proposed an attention-based scale sequence fusion algorithm to enhance the network's capability to extract multi-scale data. Shi et al. [14] used deformable convolutions to guide feature fusion blocks, promoting effective integration of multi-scale features through cross-layer and cross-scale interactions. Lin et al. [15] constructed a multi-scale feature aggregation network based on focusing and distribution mechanisms, utilizing features obtained from the backbone for efficient information exchange. Zhang [16] adopted a three-layer PAFPN structure combined with large-size feature maps to enhance the detection of small targets.

Aerial targets are often small and densely distributed, increasing the difficulty of detection. Most researchers have focused on improving feature extraction capabilities for aerial targets by enhancing the Backbone and Head components. Zhao and Zhu [17] combined Swin Transformer with SPPFS to gather global information and enhance feature information exchange, improving detection of dense objects. Sui et al. [18] introduced a dynamic detection head incorporating self-attention, which combines scale, spatial, and task-aware features to improve small target detection performance. Min et al. [19] utilized a context transformer framework to integrate global residuals and local features for detecting minute objects. Ma et al. [20] designed the Dense_CSPDarknet53 backbone network to extract latent image information. Additionally, Min et al. [21] proposed the PixED Head, which includes pixel encoders and decoders for flexible feature extraction, and used an Aux Head for online distillation to enhance feature representation.

Existing methods fall short of satisfactory results when dealing with issues such as multi-scale variations in drone imagery, small target sizes, and dense target distributions. To address these challenges, we propose a series of object detection algorithms specifically tailored to the characteristics of drone imagery. Our proposed DMFF-YOLO network introduces width and depth coefficients and can be divided into five types of networks, namely DMFF-YOLO (nano, small, middle, large, and extra), meeting different performance requirements for drone detection accuracy. The main contributions are as follows:

1) We use the RFCBAMConv module in the backbone network to interact with receptive field features to mitigate information discrepancies caused by positional changes, thereby improving the extraction of aerial target features.
2) We design the DMC and DSSFF modules in the neck network to fuse the multiscale feature maps extracted from the backbone, solving the problem of target scale changes. We added a structure specifically for detecting small target features and removed the structure features of large targets to achieve a lightweight effect.
3) We utilize Soft-NMS-CIoU, a modified strategy for determining overlapping anchor boxes, to address the issue of adjacent box suppression caused by dense targets.
4) Extensive experiments on three public datasets show that our method significantly improves the performance on aerial images primarily containing small targets. Comparisons with other advanced algorithms demonstrate the superiority and versatility of DMFF-YOLO, providing a valuable reference for related research.

II. METHODOLOGY

The overall framework of the DMFF-YOLO detection algorithm consists of three parts: Backbone, DMFF, and Head, as shown in FIGURE 1. In Backbone, we employ RFCBAMConv [22] modules to enhance feature representation within the same target. We design the DMFF structure to fuse features from continuous scales, where the DMC module concatenates features from different scales of the backbone network, and the DSSFF module utilizes semantic information from the backbone network's feature map to guide detailed information and address the issue of target scale variations. The Soft-NMS-CIoU [23], [24] changes the strategy for determining overlapping anchor boxes, improving the model's performance in detecting small targets.

A. RFCBAMConv MODULE

In standard convolution operations, the sliding window of the shared-parameter convolution kernel extracts feature information, overcoming the issue of large computational overhead in traditional fully connected layers. However, it struggles to capture the differences in information at various positions, resulting in a limited extraction of features. The attention mechanism can enhance the importance of each feature in the input feature map. By integrating the Spatial attention mechanism into the standard convolution, the limitations of parameter sharing in the convolution process can be mitigated to a certain extent, thereby improving the performance of the convolutional neural network. The mathematical calculations for standard convolution and convolution

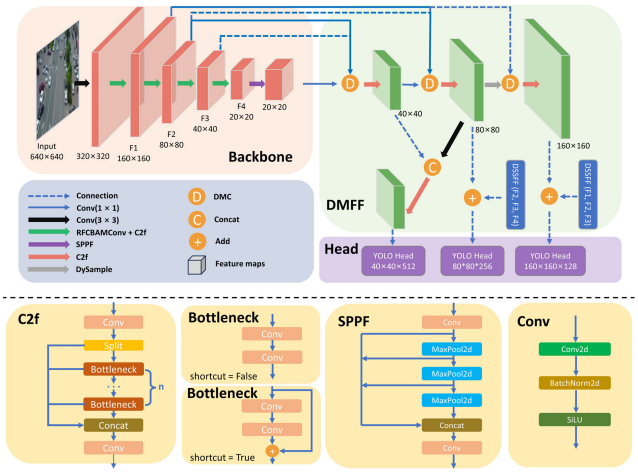


FIGURE 1. Structure of the DMFF-YOLO network.

integrated with the attention mechanism can be represented as follows:

$$F_N = X_{N1} \times K_1 + X_{N2} \times K_2 + \ldots + X_{NS} \times K_S \quad (1)$$

$$F_{AN} = X_{N1} \times A_{N1} \times K_1 + X_{N2} \times A_{N2} \times K_2 + \ldots + X_{NS} \times A_{NS} \times K_S \quad (2)$$

Here, F_N represents the output value after the standard convolution operation, F_{AN} represents the output value of introducing the spatial attention mechanism in the convolution process, K represents the convolution kernel, S represents the number of parameters in the convolution kernel, N is the total number of convolution kernels, K_S represents the s-th shared convolution parameter in the kernel, and X_N and A_N represent the values of the input feature map and attention map at different positions. However, upon careful analysis, it can be found that when using spatial attention, the sliders of each attention feature map overlap, such as $A_{12} = A_{21}, A_{13} = A_{22} \ldots$. In large 3×3 convolutions, the problem of parameter sharing is still not resolved, limiting the effectiveness of spatial attention.

In response to these questions, We introduce the RFCBAMConv convolution kernel to replace the standard convolution kernel in the main network as shown in FIGURE 2. It emphasizes the importance of different features in the receptive field slider. It prioritizes the receptive field features, thereby solving the problem of shared sliders in the convolution process being insensitive to information differences. The structure of RFCBAMConv is shown in the figure, divided into two branches, upper and lower. The upper branch uses 3×3 group convolution to extract the receptive field features of the input feature map, mapping the original features to receptive field features. It emphasizes the importance of different features in the receptive field slider through normalization and the ReLU activation function, then adjusts the shape to obtain non-overlapping receptive field features F_{rf}. After adjusting the shape, average pooling and maximum pooling on the channel dimension are performed to obtain channel distribution information. The channel attention weight ω_1 is obtained through standard convolution and sigmoid activation functions. In the lower branch, AvgPool is used to aggregate the global information of the input features. It enters two fully connected layers and uses softmax to emphasize the importance of each feature in the receptive field features, obtaining spatial attention weight ω_2. The calculation formula for the output features of RFCBAMConv is:

$$F = \omega_1 \times \omega_2 \times F_{rf} \quad (3)$$

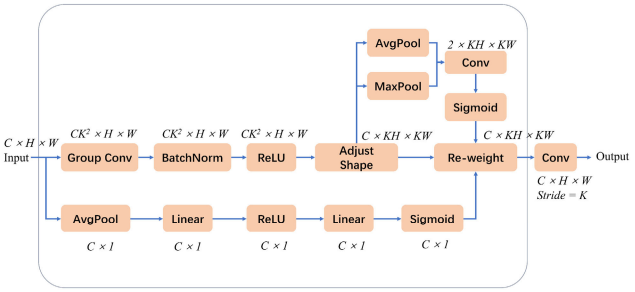


FIGURE 2. The RFCBAMConv structure.

B. DYNAMIC MULTI-SCALE FEATURE FUSION STRUCTURE

As shown in FIGURE 3, the neck structure is a DMFF structure. This structure is composed of a top-down branch and a bottom-up branch. In the top-down branch, different layers receive features of different scales through different feature maps in the main layer, namely the output feature map F5 of SPPF, and the feature maps F1, F2, F3, and F4 at different stages of the main network. Two DSSFF modules receive features of three scale sequences. The output of the bottom-up branch consists of three feature maps of different sizes, namely feature maps P2, P3, and P4, which are output to the head structure.

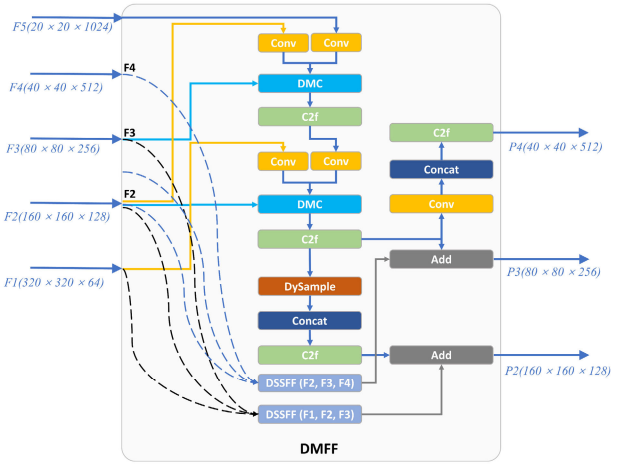


FIGURE 3. The overall architecture of DMFF.

The traditional FPN fusion mechanism only upsamples small-sized detection maps, which overlooks the rich detail information in large-scale feature layers. This results in poor detection performance for small targets that rely on detailed information, while requiring large computational overheads. We propose the DMC module, which combines three different scale feature maps in the main network, fuses the detail information in the feature map, and balances the spatial information in the lower layers with the semantic information in the higher layers. As shown in FIGURE 4, we first downsample the large-scale feature map through maximum pooling and average pooling to enrich the features of small aerial targets. For small-sized feature maps, we use DySample as the upsampling method. The feature map and the set upsampling factors are first passed through a linear layer, then reshaped into $2g \times sh \times sw$ through a pixel shuffle method, and finally the upsampling feature map $c \times sh \times sw$ is obtained through the offset. As it utilizes point-based sampling methods and learning sampling angles for upsampling, it completely avoids time-consuming dynamic convolution operations and additional subnets, improving model performance at minimal computational cost.

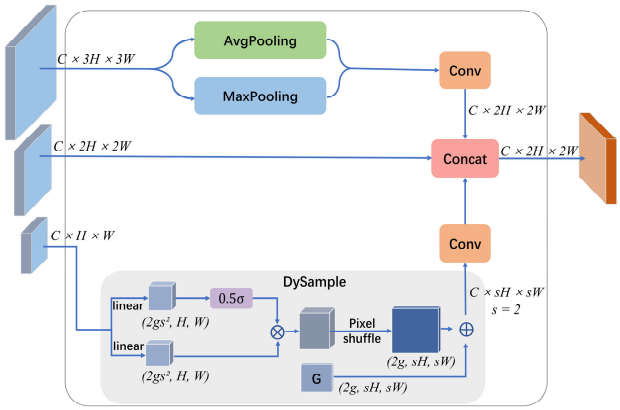


FIGURE 4. The DMC structure. The input consists of features with three scales of continuous variation, and the output is the features of the intermediate scale.

Regarding the multi-scale problem of aerial targets, the existing method is directly using the features obtained from the Neck part for target recognition. This structure cannot effectively utilize the correlation mapped by the pyramid structure of the main network. The size of the image changes during the downsampling in the main network, blurring the image's details but retaining the target's structural features. This paper designs a DSSFF module that dynamically selects the context on the feature map to enhance the resolution of the feature map, retains the target scale features after upsampling, and then uses the scale axis to build the scale space. This represents the range of various scales that an object can have. Finally, three-dimensional convolution is used to extract the scale sequence features after the three scales are stacked. As shown in FIGURE 5, the two smaller-scale feature maps are first adjusted to 256 channels using a 1×1 convolution, then the two feature maps are resized to the largest scale using DySample at 2x and 4x, respectively. The unsqueeze method adjusts each feature map to a four-dimensional tensor and stacks them into a scale sequence space along the depth dimension. Finally, three-dimensional convolution, normalization, and SiLU activation functions are used to extract the scale sequence features.

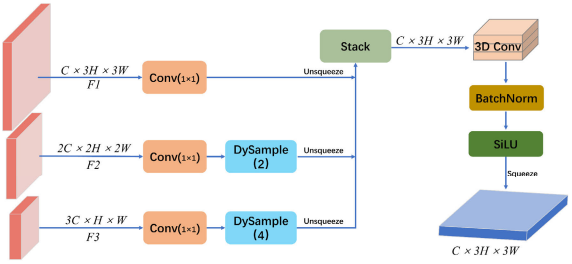


FIGURE 5. The DSSFF structure.

Most of the targets for aerial detection present small-size characteristics. The traditional YOLOv8 neck structure, the PAFPN structure, has difficulty extracting features of small targets and has many parameters. Therefore, we have redesigned the neck structure, discarding the large target fusion features and detection structures that are not suitable for small target datasets, adding a small target detection head, and using the DSSFF structure to assist in outputting features,improving the accuracy of small targets while solving the scale change issue. Experimental analysis shows that deleting the detection head for large targets will significantly reduce the number of parameters without affecting detection accuracy. Therefore, we only output the feature maps of three different sizes to three different detection heads to detect tiny, small, and medium-sized targets. The DMFF output framework is shown in FIGURE 6.

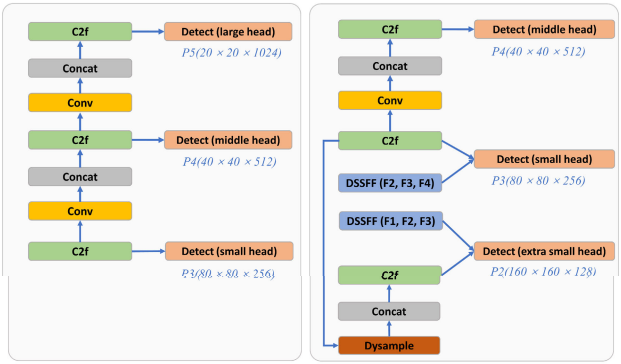


FIGURE 6. The DMFF output features with the detection head. (a) shows the feature structure and detection head of yolov8. (b) shows the feature structure and detection head of DMFF-YOLO.

C. SOFT-NMS-CIoU

Traditional Non-Maximum Suppression (NMS) can mistakenly eliminate bounding boxes that detect different objects but are close in distance. Soft-NMS works differently

from NMS. Soft-NMS calculates the Intersection-over-Union (IoU) of the highest scoring box and the bounding box, using a Gaussian function as the weight function. It lowers the score of the predicted boundary to replace the original score rather than directly setting the scores for other boxes to zero for deletion. The formulae for Soft-NMS are shown in equations (4) and (5).

$$s_i = \begin{cases} s_i, & iou(m, b_i) < N_t \\ s_i(1 - iou(m, b_i)), & iou(m, b_i) \geq N_t \end{cases} \tag{4}$$

$$s_i = s_i e^{-\frac{iou(m,b_i)^2}{\sigma}}, \forall b_i \notin D \tag{5}$$

In these formulas, m represents the linear decay of scores, meaning that detection boxes farther away will not be significantly affected, while detection boxes closer will not receive a substantial penalty. Since overlaps are not continuous, a Gaussian function addresses the discontinuity in penalties.

It should be noted that the IoU loss function is commonly used to measure the overlap between predicted boxes and ground truth boxes. However, in cases where the predicted box and the ground truth box do not intersect, the IoU loss function fails to reflect the distance between the two boxes, leading to gradient vanishing issues. The DIoU loss function considers minimizing the distance between the center points of the predicted box and the target, solving the non-intersecting loss stagnation problem and enabling faster and more stable regression. CIoU, based on DIoU, additionally accounts for the aspect ratio consistency. Therefore, we propose an improved NMS method that combines CIoU with Soft-NMS to replace IoU. The formula for calculating CIoU loss is shown in equation (6).

$$L_{CIoU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon \tag{6}$$

In the formula, c represents the diagonal length of the minimum bounding box covering two boxes, and ρ denotes the Euclidean distance between the predicted box and the ground truth box. α is a positive balance parameter, while represents the consistency of the aspect ratio, as shown in equations (7) and (8). The overlapping area factor is given a higher priority when there is no overlap.

$$\upsilon = \frac{4}{\pi^2}(arctan\frac{\omega^{gt}}{h^{gt}} - arctan\frac{\omega}{h})^2 \tag{7}$$

$$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon} \tag{8}$$

To prove the effectiveness of the proposed method, this paper compares Soft-NMS with DIoU, GIoU, EIoU, and other loss functions. The results show that Soft-NMS-CIoU outperforms the other methods.

III. EXPERIMENTS

A. DATASET DESCRIPTION AND EVALUATION METRICS

The VisDrone2019 dataset [25], collected and released by the Machine Learning and Data Mining Laboratory at Tianjin University, consists of 8629 images captured at various locations and heights. The dataset is divided into 10 categories, with 6471 images for training, 548 for validation, and 1610 for testing.

The UAVDT dataset [26] consists of 50 videos containing 40736 images, divided into 3 categories: cars, trucks, and buses. Among these, 24778 images are used for training, and 15598 images are used for testing. Similar to the segmentation dataset used in SMFF-YOLO [27], the training and testing sets are obtained from different videos, each appearing only in one dataset.

The DOTA dataset [28] is an aerial remote sensing dataset with 188,282 manually annotated instances. The training set contains 15,749 images, and the validation set contains 5,297 images. The DOTA dataset has 15 types of objects and includes challenging and complex scenes. FIGURE 7 illustrates the challenges presented by these images, necessitating a more accurate and robust detection model.

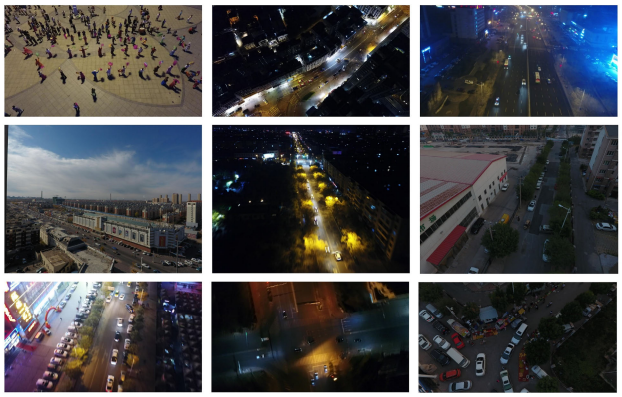


FIGURE 7. Images in the three types of datasets. (a) The challenge of small dense targets. (b) Challenges of special weather (nighttime). (c) Challenges arising from differences in occlusion and illumination.

The evaluation metrics for the experimental results include precision (P), recall (R), Average Precision and mean Average Precision (mAP) for all target categories. $mAP_{0.5}$ represents the mAP at an IoU threshold of 0.5, while $mAP_{0.95}$ indicates the average mAP across IoU thresholds from 0.5 to 0.95 with an interval of 0.05. Params represent the number of parameters in the model, and GFLOPs refer to the model's billion floating-point operations per second. Formulas (9), (10), (11) and (12) respectively denote the calculation formulas for evaluation metrics P, R, AP and mAP.

$$P = TP/(TP + FP) \tag{9}$$

$$R = TP/(TP + FN) \tag{10}$$

$$AP = \int_0^1 P(R)dR \tag{11}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{12}$$

These formulas, TP represents the number of true positive samples identified as positive by the model, FP represents

the number of false positive samples incorrectly classified as positive, and *FN* represents the number of false negative samples incorrectly classified as negative, and *N* represents the number of detection categories. These metrics are crucial for calculating precision and other evaluation metrics. All experimental results for Precision (P), Recall (R), and mAP values are reported as percentages.

B. EXPERIMENT DETAILS

In the experiment, we used Windows as the operating system, python 3.9, Pytorch 2.1.0, Cuda 12.1 as the desktop computing software environment, and the NVIDIA RTX 4060 graphics card as the hardware. The neural network code was modified based on the Ultralystic YOLOv8.1.9 version as the base model, and all hyperparameters were kept consistent during training. We provide the settings for all relevant model training parameters in the form of TABLE 1. DMFF-YOLO did not use pre-trained parameters during training. In all the experimental results listed below, all YOLOv8 and DMFF-YOLO are the results obtained from our experimental training. The remaining model results come from the relevant referenced papers.

TABLE 1. Hyperparameter settings for the model.

Hyperparameters	Value
Imgsz	640 × 640
Batch	4
Epoch	300
Weight Decay	0.0005
Momentum	0.937
Initial Learning Rate	0.01
Optimizer	SGD

C. ABLATION EXPERIMENT

To further explore the impact of different IOUs on Soft-NMS, this study conducted comparative experiments on the YOLOv8-s framework using IOU, CIOU [24], DIOU [24], EIOU [29], GIOU [30], ShapeIOU [31] methods in Soft-NMS. The experimental results are shown in TABLE 2.

TABLE 2. Comparison experiments between conventional NMS and Soft-NMS with different loss functions.

Method	P	R	$mAP_{0.5}$	$mAP_{0.95}$
NMS	53.5	42.7	45.1	27.6
Soft-NMS	62.7	37.3	50.7	34.9
Soft-NMS-DIoU	62.8	38.4	50.9	37.3
Soft-NMS-EIoU	61.6	37.9	50.3	33.8
Soft-NMS-GIoU	61.8	37.1	50.3	33.7
Soft-NMS-ShapeIoU	62.4	38.6	50.9	33.8
Soft-NMS-CIoU	62.7	38.8	51.3	38.0

Here, NMS performs poorly on small and dense datasets because it treats neighboring bounding boxes as redundant during detection. Soft-NMS effectively avoids the issue of removing redundant boxes, but IoU cannot reflect the distance between non-intersecting ground truth boxes and predicted boxes. CIoU can minimize the normalized distance between the center points of the two bounding boxes and consider the aspect ratio's consistency. Empirical evidence shows that using CIoU as the loss function for Soft-NMS achieves the best detection results.

The baseline method we used in the ablation experiment is the YOLOv8-s model, and the dataset is the VisDrone2019 dataset. The main evaluation indicators are $mAP_{0.5}$, and $mAP_{0.95}$ and parameters. The results of the ablation experiments are shown in TABLE 3.

TABLE 3. Ablation experimental results of DMFF-YOLO model.

Method	P	R	$mAP_{0.5}$	$mAP_{0.95}$	Params(M)
YOLOv8-s	51.7	38.2	39.6	23.6	11.16
+ RFCBAMConv	51.2	39.4	40.3	24.1	11.19
+ DMFF	55.8	43.1	44.7	27.2	7.65
+ Soft-NMS-CIoU	62.7	38.8	51.3	34.5	7.70

Here, it can be inferred that the overall model's $mAP_{0.5}$ evaluation index value increased by 11.7%, with a parameter reduction of 31.1%. Among the methods, Soft-NMS-CIoU has the most pronounced improvement, with a $mAP_{0.5}$ score increase of 6.6%, indicating that Soft-NMS-CIoU can effectively handle dense targets. The DMFF structure improves the $mAP_{0.5}$ by 4.4% based on RFCBAMConv, indicating that our method improves the detection effect of small targets and reduces the false detection of small targets, caused by scale change. Moreover, the number of channels in the large-scale feature map is 25% of that in the small-scale feature map, so the number of parameters is reduced by 31.1 %. When the RFCBAMConv structure is used as the convolution structure of the backbone network, the result is also improved, and the $mAP_{0.5}$ index is increased by 0.7%.

D. COMPARISON WITH THE YOLOv8 NETWORK

In this section, we comprehensively compare the DMFF-YOLO network proposed in this paper and the YOLOv8 network on the VisDrone dataset, considering detection accuracy, detection speed and computation.

From TABLE 4, it can be seen that compared to YOLOv8, our proposed DMFF-YOLO achieves better detection accuracy with fewer parameters. For instance, compared to YOLOv8-s, DMFF-YOLO (small) improves the detection accuracy by 11.7% while reducing the parameter amount by 31.1%. Although the computational load increases and the detection efficiency slows down, it achieves a detection speed of 37FPS and realizes real-time detection.

E. COMPARISON WITH STATE-OF-THE-ART METHODS

To verify the effectiveness of the improved algorithm, the experimental results of DMFF-YOLO are compared with the most advanced algorithms released on this dataset over the years. The results are shown in TABLE 5. Our DMFF-YOLO has achieved new state-of-the-art results in versions with the same parameter size. For the most

TABLE 4. Comparison between DMFF-YOLO and YOLOv8 networks.

Method	Params(M)	GFLOPS	FPS	$mAP_{0.5}$	$mAP_{0.95}$
YOLOv8-n	3.0	8.4	151	33.4	19.2
DMFF-YOLO (nano)	2.1	11.9	43	44.6	29.6
YOLOv8-s	11.1	28.5	128	39.6	23.6
DMFF-YOLO (small)	7.7	35.2	37	51.3	34.5
YOLOv8-m	25.8	79.3	91	43.5	26.5
DMFF-YOLO (middle)	18.9	92.7	24	53.6	36.4
YOLOv8-l	43.7	165.2	72	43.6	26.9
DMFF-YOLO (large)	34.8	193.9	18	54.7	37.1
YOLOv8-x	68.2	257.8	47	46.6	28.9
DMFF-YOLO (extra)	54.1	298.8	10	55.5	37.6

advanced YOLO algorithm, YOLOv9-C, by loading pre-trained weights and adding auxiliary detection heads, the mAP50 has increased to 48.1%, and GFLOPS has increased from 102.1 to 237.8, surpassing the YOLOv8-x model, but still lower than our proposed algorithm. This indicates that our method has demonstrated excellent detection accuracy on the technical difficulties unique to drone images, such as small target size, target scale changes, and dense target distribution, to some extent overcoming the challenges of aerial images.

TABLE 5. Results of different algorithms are compared on the VisDrone2019 dataset, and the best result is shown in bold.

Method	$mAP_{0.5}$	$mAP_{0.95}$	Params(M)	GFLOPS
Faster R-CNN [7]	35.6	19.4	41.1	168.25
YOLOv5-s [32]	33.3	17.9	7.2	16.5
YOLOv8-s [33]	39.6	23.6	11.1	28.5
YOLOv9-C [34]	48.1	29.9	50.9	237.8
BDH-YOLO [18]	42.9	26.2	9.39	-
Drone-YOLO (large) [16]	51.3	31.9	76.2	-
PVswin-YOLOv8-s [35]	43.3	26.4	21.6	-
YOLO-DCTI [19]	49.8	27.4	37.6	-
DMFF-YOLO (small)	51.3	34.5	7.7	35.2

We conducted Experiments on the UAVDT and DOTA datasets to demonstrate the model's generalization ability, as shown in TABLE 6 and TABLE 7. Compared to the current methods, DMFF-YOLO shows a significant improvement in the mAP, $mAP_{0.5}$, and $mAP_{0.75}$ indices on the UAVDT dataset, with a performance improvement of at least 3% higher than other advanced methods. The DOTA dataset primarily features small and dense targets, and does not exhibit significant multi-scale issues compared to aerial datasets. As a result, the performance gap compared to mainstream models is relatively small. However, our method achieves the lowest number of parameters while maintaining comparable detection accuracy. In conclusion, this method shows good detection accuracy and a lower parameter amount for small targets and dense scenes.

TABLE 6. Comparison of results between our method and mainstream methods on UAVDT dataset.

Method	$mAP_{0.5}$	$mAP_{0.75}$	$mAP_{0.95}$
YOLOv8-s [33]	32.8	22.6	20.1
GLSAN [36]	30.5	21.7	19.0
ClusDet [37]	26.5	12.5	13.7
DMNet [38]	24.6	16.3	14.7
GDFNet [39]	26.1	17.0	15.4
SODNet [40]	29.9	18.0	17.1
DSHNet [41]	30.4	19.7	17.8
CDMNet [42]	35.5	22.4	20.7
UFPMP-Net [43]	38.7	28.0	24.6
DMFF-YOLO (small)	41.1	32.1	27.5

TABLE 7. Comparison of results between our method and mainstream methods on DOTA dataset.

Method	P	R	Params(M)	$mAP_{0.5}$	$mAP_{0.95}$
YOLOv5l	75.8	69.9	46.2	71.8	47.3
YOLOv7-tiny	77.0	66.4	13.3	69.8	44.7
FCOS[44]	75.3	69.7	23.5	72.1	46.5
ATSS[45]	76.8	76.0	42.16	73.3	47.8
YOLOv8-S	78.3	66.2	11.2	71.6	48.4
DMFF-YOLO (small)	76.4	67.6	7.7	72.1	48.9

To demonstrate the competitiveness of DMFF-YOLO compared to other lightweight models, we compared DMFF-YOLO (nano) with several representative lightweight algorithms on the VisDrone dataset, including GCL-YOLO [46] and PP-PicoDet [47], as shown in TABLE 8.

TABLE 8. Comparison of DMFF-YOLOv8(nano) with classical lightweight networks on the VisDrone2019 dataset with the best results in bold.

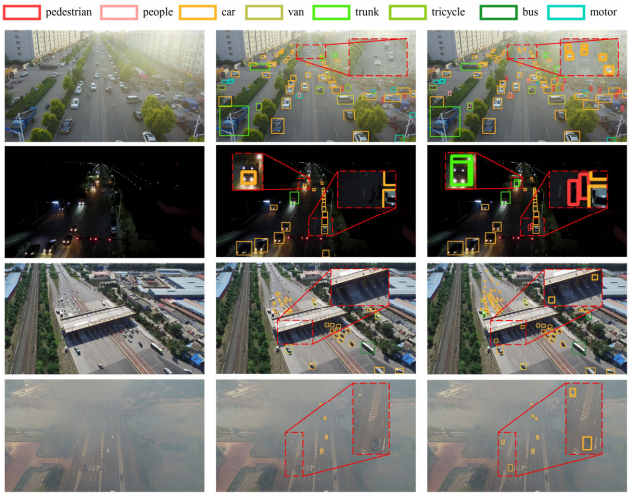
NetWork	Params(M)	GFOLPS	$mAP_{0.5}$	AP(%)								
				Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning Tricycle	Bus	Motor
YOLOv3-Tiny	8.68	12.9	15.9	19.4	18.4	3.2	49.9	12.7	9.7	8.2	4.0	14.6	18.9
YOLOv4-Tiny	5.89	7.0	19.5	21.1	25.1	4.3	57.5	15.1	15.0	12.0	5.3	23.7	16.4
YOLOv5-Lite-G	5.39	15.2	27.3	34.6	26.6	7.7	69.3	28.4	24.0	13.8	6.7	28.3	33.4
Nanodet-Plus-M-1.5x	2.44	**3.0**	30.4	27.9	24.1	7.4	73.0	35.1	27.8	17.9	8.4	49.3	33.3
YOLOv5-S	7.04	15.8	32.7	38.9	31.6	11.3	72.4	34.8	28.5	19.3	9.5	43.4	37.0
YOLOv7-Tiny	6.03	13.1	36.8	41.5	38.3	11.9	77.3	38.9	29.1	23.4	11.7	48.6	47.1
YOLOv8-m	25.8	78.7	43.4	47.0	36.8	17.4	**81.9**	49.0	**42.7**	32.0	18.6	**59.9**	48.9
GCL-YOLO-S	**1.64**	10.7	39.6	48.0	37.6	15.7	81.1	42.5	32.7	26.4	12.4	53.8	46.0
YOLOX-Tiny	5.04	15.3	31.3	35.8	21.9	9.6	73.3	34.7	28.1	18.1	10.2	46.3	34.9
PP-PicoDet-L	3.30	8.9	34.2	40.2	35.3	12.8	75.6	35.4	29.3	21.1	12.1	44.3	36.3
DMFF-YOLO(nano)	2.10	11.9	**44.6**	**50.1**	**43.4**	**21.7**	80.3	**49.5**	38.9	**34.5**	**19.8**	58.3	**49.8**

It shows that the proposed DMFF-YOLO (nano) achieved a better accuracy compared to other lightweight models with few parameters. Compared to YOLOv8-m, DMFF-YOLO (nano) reduces parameters by 91.8% and computation by 84.8% while improving detection accuracy by 1.2%. This further demonstrates that the proposed method's lightest version performs better.

F. VISUALIZATION

To further validate the effectiveness of our model, we conducted a comparative analysis between the DMFF-YOLO (small) model and YOLOv8-s. We extracted partial data from the VisDrone2019 and UAVDT datasets and compared the prediction results, as shown in FIGURE 8. In the figure, different bounding boxes correspond to different detected objects: yellow boxes represent cars, green boxes represent

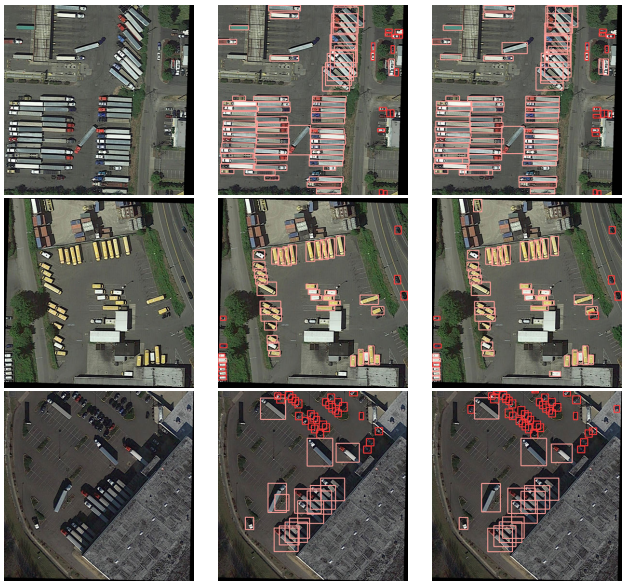
trucks, and red boxes represent pedestrians. The red dashed boxes indicate the zoomed-in images.



(a) original image (b) YOLOv8-S (c) DMFF-YOLO

FIGURE 8. Detection results of different YOLO-based methods in Visdrone and UAVDT datasets.

From the comparison results in the first row of FIGURE 8, it can be seen that DMFF-YOLO can detect smaller objects. According to the comparison images in the second row, it is found that in dimly lit conditions at night, YOLOv8 struggles to effectively detect pedestrians in low light, while DMFF-YOLO can detect pedestrians next to vehicles, and the second green box also correctly detects a truck. In the third image, YOLOv8 often fails to detect occluded objects, while DMFF-YOLO successfully identifies them. Lastly, even in special weather conditions like foggy weather, DMFF-YOLO can still detect targets. In summary, DMFF-YOLO demonstrates outstanding detection capabilities when faced with the common issues of missed detections and false positives in dense, small objects. This proves the excellent detection performance of the method proposed in this paper.



(a) original image (b) YOLOv8-S (c) DMFF-YOLO

FIGURE 9. Results of DMFF-YOLO on the DOTA dataset. Red boxes indicate detections classified as cars, while pink boxes denote detections classified as large vehicles.

To demonstrate the model's versatility, we compared it with the original model on the remote sensing dataset DOTA. As shown in FIGURE 9, in the three comparison images, DMFF-YOLO detects cars more accurately, addressing the issue of YOLOv8-S struggling with small targets and missing detections. This highlights the improved model's significant advantage in recognizing minute targets.

IV. CONCLUSION

In this study, we propose the DMFF-YOLO algorithm for aerial object detection based on dynamic multiscale feature fusion. According to the target characteristics of aerial images, we improve the model's feature extraction, feature fusion, and post-processing stages. We utilize the RFCBAMconv structure to address the limited feature extraction issue. For feature fusion, we introduce the DMFF structure, which merges feature maps at continuous scales in the spatial dimension of the backbone network, resolving the multiscale problem in aerial images. We incorporate feature structures and detection heads specifically designed for small objects while subtracting those for large objects, improving accuracy while reducing parameter count. Lastly, to address the common characteristics of dense and overlapping small objects, we replace the traditional NMS method with Soft-NMS-CIoU, resolving the issue of type determination for overlapping objects in dense scenarios. Experiments demonstrate that our method significantly improves the challenging issue of detecting UAV targets.

While our series of methods cater to the needs of different devices, there are still some limitations. In subsequent work, we may use a lightweight backbone network or pruning methods to further lighten the aerial detection model. Our focus going forward is to achieve a better lightweight model while maintaining the model's excellent detection performance.

REFERENCES

[1] L. P. Osco, J. M. Junior, A. P. M. Ramos, L. A. D. C. Jorge, S. N. Fatholahi, J. D. A. Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li, "A review on deep learning in UAV remote sensing," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102456.

[2] Y. Yu, T. Gu, H. Guan, D. Li, and S. Jin, "Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1894–1898, Dec. 2019.

[3] H. Pu, X. Chen, Y. Yang, R. Tang, J. Luo, Y. Wang, and J. Mu, "Tassel-YOLO: A new high-precision and real-time method for maize tassel detection and counting based on UAV aerial images," *Drones*, vol. 7, no. 8, p. 492, Jul. 2023.

[4] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-UAV," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 82–89, Apr. 2018.

[5] J. Liao, Y. Piao, J. Su, G. Cai, X. Huang, L. Chen, Z. Huang, and Y. Wu, "Unsupervised cluster guided object detection in aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11204–11216, 2021.

[6] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

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

[9] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.

[10] F. Li, A. Zeng, S. Liu, H. Zhang, H. Li, L. Zhang, and L. M. Ni, "Lite DETR : An interleaved multi-scale encoder for efficient DETR," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18558–18567. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Li_Lite_DETR_An_Interleaved_Multi-Scale_Encoder_for_Efficient_DETR_CVPR_2023_paper.html

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," presented at the *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021. Accessed: Jul. 27, 2024. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper

[12] G. Jocher, A. Stoken, and J. Borovec. (Jun. 2020). *v7.0—YOLOv5 SOTA Realtime Instance Segmentation*. [Online]. Available: https://github.com/ultralytics/yolov5/releases/tag/v7.0

[13] M. Kang, C.-M. Ting, F. Fung Ting, and R. C.-W. Phan, "ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation," 2023, *arXiv:2312.06458*.

[14] Y. Shi, C. Wang, S. Xu, M.-D. Yuan, F. Liu, and L. Zhang, "Deformable convolution-guided multiscale feature learning and fusion for UAV object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3362890.

[15] Y. Lin, J. Li, S. Shen, H. Wang, and H. Zhou, "GDRS-YOLO: More efficient multiscale features fusion object detector for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3397717.

[16] Z. Zhang, "Drone-YOLO: An efficient neural network method for target detection in drone images," *Drones*, vol. 7, no. 8, p. 526, Aug. 2023.

[17] L. Zhao and M. Zhu, "MS-YOLOv7: YOLOv7 based on multi-scale for object detection on UAV aerial photography," *Drones*, vol. 7, no. 3, p. 188, Mar. 2023.

[18] J. Sui, D. Chen, X. Zheng, and H. Wang, "A new algorithm for small target detection from the perspective of unmanned aerial vehicles," *IEEE Access*, vol. 12, pp. 29690–29697, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10433537/

[19] L. Min, Z. Fan, Q. Lv, M. Reda, L. Shen, and B. Wang, "YOLO-DCTI: Small object detection in remote sensing base on contextual transformer enhancement," *Remote Sens.*, vol. 15, no. 16, p. 3970, Aug. 2023.

[20] C. Ma, Y. Fu, D. Wang, R. Guo, X. Zhao, and J. Fang, "YOLO-UAV: Object detection method of unmanned aerial vehicle imagery based on efficient multi-scale feature fusion," *IEEE Access*, vol. 11, pp. 126857–126878, 2023, doi: 10.1109/ACCESS.2023.3329713.

[21] X. Min, W. Zhou, R. Hu, Y. Wu, Y. Pang, and J. Yi, "LWUAVDet: A lightweight UAV object detection network on edge devices," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 24013–24023, Jul. 2024, doi: 10.1109/JIOT.2024.3388045.

[22] X. Zhang, C. Liu, D. Yang, T. Song, Y. Ye, K. Li, and Y. Song, "RFAConv: Innovating spatial attention and standard convolutional operation," 2023, *arXiv:2304.03198*.

[23] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5562–5570. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2017/html/Bodla_Soft-NMS_–_Improving_ICCV_2017_paper.html

[24] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000. Accessed: Apr. 19, 2024. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/6999

[25] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, L. Bo, and H. Shi, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 213–226. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_ICCVW_2019/html/VISDrone/Du_VisDrone-DET2019_The_Vision_Meets_Drone_Object_Detection_in_Image_Challenge_ICCVW_2019_paper.html

[26] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 370–386. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_ECCV_2018/html/Dawei_Du_The_Unmanned_Aerial_ECCV_2018_paper.html

[27] Y. Wang, H. Zou, M. Yin, and X. Zhang, "SMFF-YOLO: A scale-adaptive Yolo algorithm with multi-level feature fusion for object detection in UAV scenes," *Remote Sens.*, vol. 15, no. 18, p. 4580, Sep. 2023.

[28] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983. Accessed: Jun. 20, 2024. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Xia_DOTA_A_Large-Scale_CVPR_2018_paper.html

[29] Z. Yang, X. Wang, and J. Li, "EIoU: An improved vehicle detection algorithm based on VehicleNet neural network," *J. Phys., Conf.*, vol. 1924, no. 1, May 2021, Art. no. 012001. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1924/1/012001/meta

[30] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 658–666. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Rezatofighi_Generalized_Intersection_Over_Union_A_Metric_and_a_Loss_for_CVPR_2019_paper.html

[31] H. Zhang and S. Zhang, "Shape-IoU: More accurate metric considering bounding box shape and scale," 2023, *arXiv:2312.17663*.

[32] *YOLOv5*. Accessed: 15, Mar. 2023. [Online]. Available: https://github.com/ultralytics/yolov5

[33] G. Jocher. *Ultralytics YOLOv8: V6*. Accessed: Oct. 23, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[34] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.

[35] N. U. A. Tahir, Z. Long, Z. Zhang, M. Asim, and M. El Affendi, "PVswin-YOLOv8s: UAV-based pedestrian and vehicle detection for traffic management in smart cities using improved YOLOv8," *Drones*, vol. 8, no. 3, p. 84, Feb. 2024.

[36] S. Deng, S. Li, K. Xie, W. Song, X. Liao, A. Hao, and H. Qin, "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2021.

[37] 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. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2019/html/Yang_Clustered_Object_Detection_in_Aerial_Images_ICCV_2019_paper.html

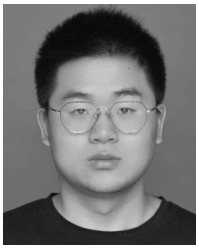
[38] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 737–746. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content_CVPRW_2020/html/w11/Li_Density_Map_Guided_Object_Detection_in_Aerial_Images_CVPRW_2020_paper.html

[39] R. Zhang, Z. Shao, X. Huang, J. Wang, and D. Li, "Object detection in UAV images via global density fused convolutional network," *Remote Sens.*, vol. 12, no. 19, p. 3140, Sep. 2020.

[40] G. Qi, Y. Zhang, K. Wang, N. Mazur, Y. Liu, and D. Malaviya, "Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion," *Remote Sens.*, vol. 14, no. 2, p. 420, Jan. 2022.
[41] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3257–3266. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content/WACV2021/html/Yu_Towards_Resolving_the_Challenge_of_Long-Tail_Distribution_in_UAV_Images_WACV_2021_paper.html
[42] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2789–2798. Accessed: Apr. 19, 2024. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2021W/VisDrone/html/Duan_Coarse-Grained_Density_Map_Guided_Object_Detection_in_Aerial_Images_ICCVW_2021_paper.html
[43] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1026–1033. Accessed: Apr. 19, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19986
[44] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.
[45] L. J. Biffi, E. Mitishita, V. Liesenberg, A. A. D. Santos, D. N. Gonçalves, N. V. Estrabis, J. D. A. Silva, L. P. Osco, A. P. M. Ramos, J. A. S. Centeno, M. B. Schimalski, L. Rufato, S. L. R. Neto, J. M. Junior, and W. N. Gonçalves, "ATSS deep learning-based approach to detect apple fruits," *Remote Sens.*, vol. 13, no. 1, p. 54, Dec. 2020.
[46] J. Cao, W. Bao, H. Shang, M. Yuan, and Q. Cheng, "GCL-YOLO: A GhostConv-based lightweight Yolo network for UAV small object detection," *Remote Sens.*, vol. 15, no. 20, p. 4932, Oct. 2023.
[47] G. Yu, Q. Chang, W. Lv, C. Xu, C. Cui, W. Ji, Q. Dang, K. Deng, G. Wang, Y. Du, B. Lai, Q. Liu, X. Hu, D. Yu, and Y. Ma, "PP-PicoDet: A better real-time object detector on mobile devices," 2021, *arXiv:2111.00902*.

XIAOYANG QIU received the Bachelor of Engineering degree from Chongqing University of Posts and Telecommunications, in 2023. He is currently pursuing the master's degree in engineering with China West Normal University. His research interests include computer vision, target tracking, and object detection.

YAJUN CHEN received the master's degree in radio electronics from Central China Normal University, in 1993. He is currently a Professor with China West Normal University, Nanchong, China. His research interests include artificial intelligence, intelligent information processing, and embedded systems. He has published approximately 70 research articles in these areas.



CHAOYUE SUN received the bachelor's degree from Hebei University of Engineering, in 2022. He is currently pursuing the master's degree in engineering with China West Normal University. His research interests include computer vision and object detection. He has published several articles and participated in many research projects in this field.

JIANYING LI received the bachelor's degree from Chongqing University of Technology, in 2023. She is currently pursuing the master's degree in engineering with China West Normal University. Her research interests include computer vision and object detection.

MEIQI NIU received the bachelor's degree from the North University of China, in 2023. She is currently pursuing the master's degree in engineering with China West Normal University. Her research interests include object detection and computer vision.