DUCAF-Net : An Object Detection Method for UAV Imagery

Yuhang Bai, Zhengpeng Li, Member, IANEG, Jiansheng Wu*, and Xinmiao Yu

Abstract—**This paper proposes an object detection method called "Down Up Coordinate Attention Fusion model (DUCAF-Net)" to address the challenges of object density and complex backgrounds in aerial images captured by drones. DUCAF-Net integrates coordinate attention at different resolutions, aiming to learn spatial coordinate information from feature maps at various resolutions, enhancing the expression capability of spatial features, and simultaneously reducing the diffusion of features from small dense objects and the phenomenon of feature coupling. DUCAF-Net introduces deformable convolutions and transposed convolutions, and designs an upsampling module to increase the receptive field of feature maps, better capturing the details of target features, thus improving the object detection performance. The AP score on the VisDrone2019 test set is 22.9. DUCAF-Net demonstrates satisfactory performance in medium-scale object detection and also performs well in small-scale object detection. The experimental results show that DUCAF-Net's performance is delightful.**

Index Terms—**deep learning, Aerial images object detection, Attention mechanism, Convolutional neural networks (CNNs)**

I. Introduction

In recent years, the development of unmanned aerial vehicle (UAV) aerial photography technology has led to an increasing number of high-quality images emerging. Aerial images have significant applications, including environmental monitoring, UAV-assisted maritime rescue, resource surveying, and more.Object detection is a crucial research area in computer vision, finding extensive applications across various domains. Target detection in aerial images captured by unmanned aerial vehicles (UAVs) finds wide application in tasks like military operations, agriculture, and rescue missions. Nevertheless, UAV aerial datasets present several challenges, including larger image resolutions and more complex backgrounds. Therefore, directly applying classical deep learning algorithms to accomplish the task of object detection is not feasible[1].

An increasing number of researchers are delving into UAV object detection through deep learning. Several methods have been introduced in this domain to cater to the unique traits of aerial datasets. For example, in terms of network design, researchers utilize deeper networks and incorporate attention mechanisms to boost detection accuracy and coverage. In data augmentation, various techniques are employed by researchers to enrich data, tackling challenges like intricate backgrounds and sample imbalances in aerial datasets[2]. These innovative approaches consistently enhance the efficiency and performance of aerial image object detection, bolstering its application and evolution.

This paper proposes the down up coordinate attention fusion model (DUCAF-Net) for target detection. The goal is to improve UAV image object detection performance, especially in scenarios with dense targets and intricate backgrounds.

DUCAF-Net incorporates the coordinate attention mechanism and introduces the multi-resolution coordinate attention fusion module (MCAF). This module integrates coordinate offset weights from various resolutions and dimensions, dynamically capturing the aspect ratios of targets. The module not only efficiently extracts feature maps across multiple scales but also seamlessly merges global information from low-resolution images with detailed insights from high-resolution images, significantly enhancing the network's detection accuracy.

DUCAF-Net introduces the DcUp upsampling module. This module combines deformable convolution and transposed convolution techniques, effectively enlarging feature map dimensions. The module adapts to targets of different scales and shapes. This design enables the model to capture target details with heightened sensitivity and bolsters recognition capabilities. Additionally, the module reduces feature diffusion and enhances detection accuracy and stability.

II. Related Work

A. Object Detection Methods

The task of object detection is to identify the presence of objects in an image and output their categories and locations. Based on the deep learning detection process, object detection can be categorized into two main methods: Two-Stage and One-Stage [3]. The Two-Stage method consists of two phases: region proposal and object detection. While it offers superior performance, its execution speed is slower. The One-Stage method processes both region proposals and object detection within a single network. Compared to Two-Stage, it operates faster but with slightly reduced accuracy.

Faster R-CNN [4], a representative Two-Stage object detection network, uses the region proposal network (RPN) to speed up model convergence, supplanting the traditional region proposal method. RPN predicts both detection regions and object confidence simultaneously. Its strength is

Manuscript revised May 17, 2023; revised August 30, 2023. The research work was supported by National Natural Science Foundation of China (No.51774179), and Science and Technology Innovation Project of University of Science and Technology Liaoning (LKDYC202109 and LKDYC202219).

Yuhang Bai is a graduate student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 604760451@qq.com)

Zhengpeng Li is a doctoral student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 1156361257@qq.com).

Jiansheng Wu* is a professor of University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: ssewu@163.com).

Xinmiao Yu is a graduate student of University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 2749936763@qq.com).

in sharing convolutional layers with the object detection network, reducing detection time. Conversely, Yolov3 [5] is a One-Stage network model with superior detection performance, preferring binary cross-entropy loss to Softmax loss. To enhance detection, Yolov3 integrates residual modules, employs darknet-53 as its core structure, deepens the network, and leverages a feature pyramid for multi-scale detection. However, traditional object detection struggles with small objects, often yielding unsatisfactory results.

B. Methods for Detecting Medium and Small Objects

Advances in deep learning have significantly improved the performance of small and medium object detection[6]. However, detecting small and medium objects remains challenging, especially when the objects have low contrast, are blurry, or obscured. Researchers are dedicated to developing new object detection models to enhance performance. Researchers have employed various strategies to improve object detection models, such as introducing attention mechanisms, refining feature pyramid algorithms, and optimizing loss functions. These strategies have enhanced the detection of small objects.

Liu et al [7] optimized the anchoring method and feature fusion of Yolov3 and introduced the GIou loss. Their proposed Cross-PaNet replaced the FPN in Yolov3. While it improved maritime object detection accuracy, the bounding box detection for small objects remained imprecise. Kang et al [8] introduced an alignment matching strategy to enhance the semantic output layer, considering aspect ratio and center distance, replacing the Iou matching in SSD, thereby improving small object detection. Liu et al [9] refined Yolov5, launching Yolov5-Tassel with a bidirectional feature pyramid and robust attention module. This enhanced cross-scale feature fusion, especially excelling in detecting small corn tassels. Addressing feature scale and task contradictions, Yang et al [10] and colleagues proposed a detection framework based on RetinaNet and RBox. They used scale calibration to align the main and target feature map ratios. However, this method exhibited significant feature diffusion in complex background detection. Sun et al [11] introduced the category position (CP) module, optimizing the positional regression features of fcos. They generated guiding vectors from classification features, enhancing object localization in complex scenarios. To address blurred areas during training, they redesigned classification and bounding box regression to minimize the impact of blurred regions. However, this increased the computational complexity of the model, affecting training speed.

C. Aerial Object Detection Methods

With the proliferation and application of drone technology, the number of drone aerial photography datasets continues to grow. This provides researchers in the computer vision field with a wealth of data resources. Many researchers have begun exploring how to effectively apply traditional natural image object detectors to the task of object detection in drone aerial images. Due to the specific perspectives, resolutions, and background characteristics of drone aerial images, traditional object detection algorithms need to be adjusted or optimized. Consequently, these researchers have started to apply deep learning models to drone aerial images to enhance the accuracy and efficiency of object detection.

Nehru et al [12] proposed an enhanced Yolo-based algorithm to address specific challenges of object detection in drone imagery. They conducted a thorough analysis of the traditional Yolo algorithm's limitations in drone image object detection, highlighting issues like low accuracy, slow detection speed, missed detections, and false alarms. To tackle these challenges, they implemented various innovative strategies. They applied bounding box dimension clustering to enhance the accuracy of predicting object positions and sizes. Moreover, they used pre-trained networks for classification, leveraging a large annotated dataset to boost the model's initial performance. To enhance detection capabilities across various scales, they incorporated multi-scale detection training. However, although the approach performs well in most situations, it requires further refinement for dense object detection.

Luo [13] et al developed the Yolo-Drone detection method, a novel approach tailored for object detection challenges in drone aerial imagery. The method harnesses the unique properties of activation functions, choosing distinct functions for both shallow and deep networks. For a more accurate computation of bounding box regression loss, they adopted the EIou loss. Additionally, they integrated the improved efficient channel attention (IECA) module and utilized the pyramid pooling module to bolster the model's adaptability to intricate backgrounds and detection precision. These design enhancements specifically target the challenges of detection accuracy in drone aerial imagery stemming from complex backgrounds.

Addressing the challenges posed by noise and other interferences in infrared aerial imagery with complex backgrounds, Fang et al [14] proposed a novel method. This approach shifts the focus from object detection in small drones to predicting residual images. The model is designed to learn directly from input infrared images and map them to their corresponding residual images. For a more effective capture of local and contextual structural information, they incorporated both global and local dilated residual convolution blocks, leading to a deep fusion of features.

Addressing the challenges posed by scale variability in geospatial objects and the complexity of aerial imagery backgrounds, Guo et al.[15] and his team developed the DA2FNet, a new anchor-point detection network leveraging density maps and attention mechanisms. Using image-level supervision, the network estimates probabilities, providing a comprehensive understanding of target scales. They integrated a composite attention network to boost detection accuracy, emphasizing foreground objects and achieving superior performance in many scenarios. However, the network's performance can improve when detecting occluded objects.

III. Research Method

This section presents a comprehensive overview of the method named down up coordinate attention fusion model (DUCAF-Net), proposed in this paper and illustrated in Figure 1. The data flow of this method starts by passing through the backbone network, then proceeds through the MCAF module and DcUp module, and finally goes through the detection head.

Specifically, In the MCAF module, the process starts with the selection of the feature map with the highest resolution from four feature maps with coordinate attention, the chosen

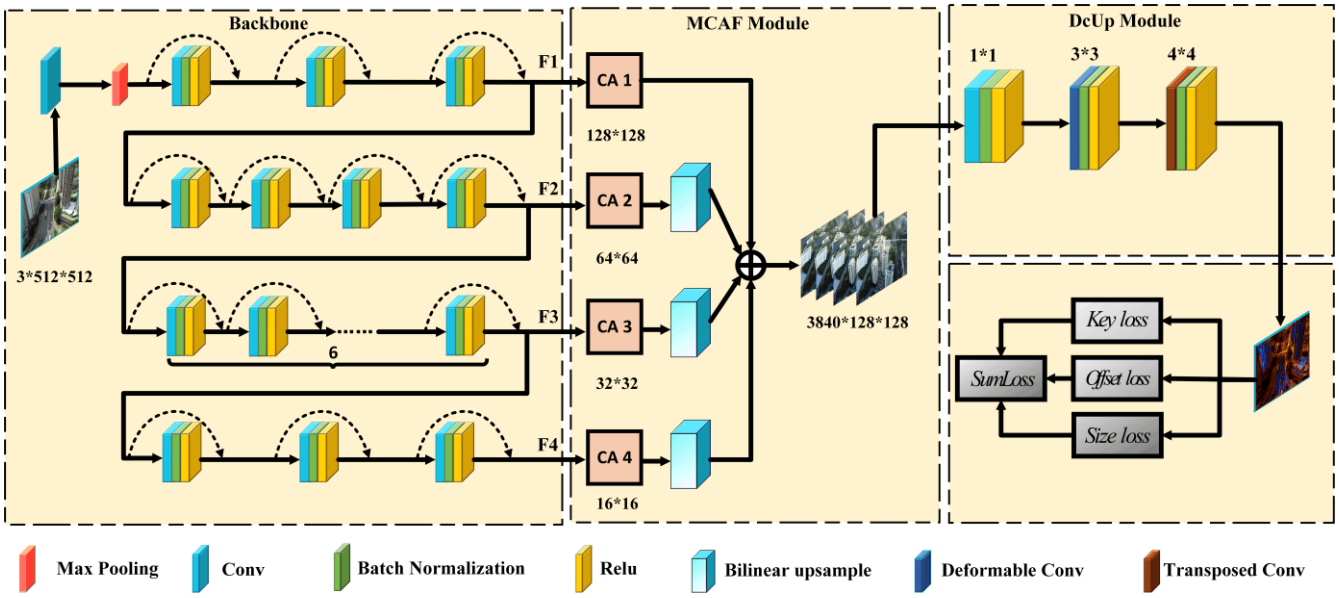


Figure 1. The framework of DUCAF-Net. ⊕ in the MCAF module represents the concatenation operation.

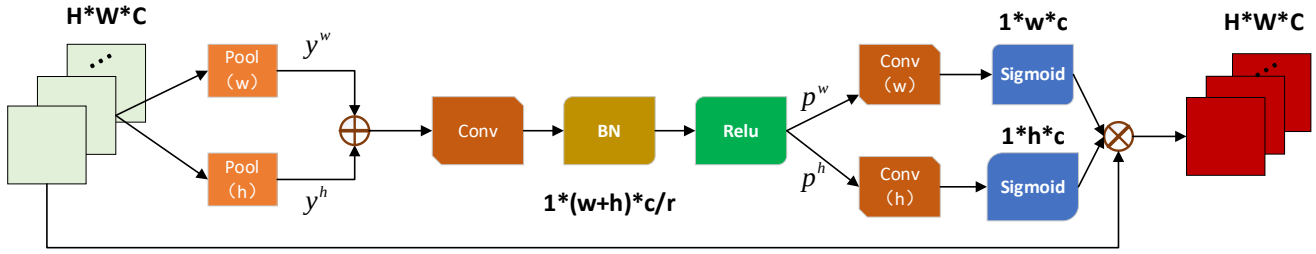


Figure 2 Coordinate attention

feature map serves as the reference. Then, the other feature maps are resized to match the size of the reference feature map, and finally, all the feature maps are fused together to obtain the fused feature map with multi-resolution coordinate attention.

The DcUp module begins by downsampling the feature map to reduce its dimension. Deformable convolution and transpose convolution are then applied to the normalized feature map to enhance the network's non-linearity and increase the receptive field of the feature map. Finally, this module outputs the resulting feature map.

In this paper, we adopt three loss functions, namely the key loss, size loss, and offset loss, to form the final loss function. The sum of the three loss values is used as the overall loss function. The process of calculating the loss will be introduced in Section 3.4.

A. *Backbone*

DUCAF-Net utilizes ResNet50 as the backbone network, employing residual learning to achieve a deeper network architecture and improve the model's performance. ResNet50 consists of four stages. In this paper, we utilize the output of each stage as the initial source for further feature enhancement, as illustrated in Figure 1. The $F1$ feature map has a size of 128×128, the $F2$ feature map has a size of 64×64, the $F3$ feature map has a size of 32×32, and the $F4$ feature map has a size of 16×16.

B. *CA Module*

Coordinate attention is an effective method that can improve object detection performance[16]. As shown in Figure 2, by incorporating spatial and channel information to adjust attention weights, coordinate attention can capture the correlation between different pixel positions, enabling the model to more accurately locate and recognize regions of interest, thus improving overall object detection performance. In this paper, we use two average pooling layers with pooling kernel sizes of $(H,1)$ and $(1,W)$, respectively. For each channel of the input feature map, position encoding is performed in both the height and width directions. This process can be referred to as formula 1 and formula 2, respectively.

$$y_c^h(h)=\frac{1}{W}\sum_{m=0}^{W}x_c(h,m), \tag{1}$$

$$y_c^w(w)=\frac{1}{H}\sum_{n=0}^{H}x_c(n,w), \tag{2}$$

where x_c is the input, y_c^h represents the encoding of the C-th channel with height h. y_c^w represents the encoding of the C-th channel with weight w.

The position encoding information for both width and height directions is concatenated and computed using a convolutional transformation function, shown in formula 3.

$$f=\delta(F_1([y^h,y^w])), \tag{3}$$

where, δ represents applying the non-linear transformation function ReLU to the normalized feature. F_1 represents the convolution operation performed on the concatenated feature map, with a kernel size of 1×1. f represents the feature map that interacts with width and height position encoding information.

Continuing, we further calculate the weights p^h and p^w in

the width and height directions respectively by using f, the original height and width of the image. Refer to formulas 4 and 5 for the calculation.

$$p^h = \sigma(F_h(f^h)), \tag{4}$$

$$p^w = \sigma(F_w(f^w)), \tag{5}$$

where, σ represents the Sigmoid activation function, f^h and f^w respectively represent the feature maps in the height and width directions, while F_h and F_w represent two convolution kernels that operate as 1x1 convolutions.

Finally, the two weights are multiplied to the original feature map, and the calculation is expressed by formula 6.

$$z_c(m,n) = x_c(m,n) \cdot p_c^h(m) \cdot p_c^w(n), \tag{6}$$

C. MCAF Module

This paper introduces a multi-resolution coordinate attention fusion method called MCAF module. The method adapts to the sampling positions of each pixel in the input image on the feature map through learning variable coordinate offsets. This enables the network to capture channel information and focus on coordinate positions in the feature map simultaneously. The module combines high-resolution and low-resolution feature maps to enhance the model's feature representation ability and alleviate feature diffusion and coupling issues in the network.

The MCAF module takes $F1$, $F2$, $F3$ and $F4$ of the backbone network as input, and utilizesthe feature map of the highest resolution $F1$ as the reference. The smaller feature maps $F2$, $F3$, and $F4$ are resized to match the size of the reference feature map. For example, when provided with the input image $F2$, the MCAF module performs horizontal and vertical interpolation to obtain the interpolated image I_{o2}, as shown in formula 7.

$$\begin{aligned} I_{o2}(i,j) = (1-\delta)\cdot(1-\beta)\cdot I_x(\tilde{m},\tilde{n}) + \delta\cdot(1-\beta)\cdot I(\tilde{m},\tilde{n}+1) \\ +(1-\delta)\cdot\beta\cdot I_x(\tilde{m}+1,\tilde{n}) + \delta\cdot\beta\cdot I(\tilde{m}+1,\tilde{n}+1), \end{aligned} \tag{7}$$

where $\tilde{m}$ denotes the integer obtained by rounding i/k down wards, $\tilde{n}$ denotes the integer obtained by rounding j/k do wnwards, $\delta = i/k - \tilde{m}$ and $\beta = j/k - \tilde{n}$ represents weights. k represents the scaling factor. k is the ratio of the input im age size to the feature map size.

Finally, the four feature maps (F_1, I_{o2}, I_{o3} and I_{o4}), each with a size of 128×128, are concatenated to form a highly expressive feature map, as shown in formula 8. The output feature map y_o has a size of $3840*128*128$.

$$y_o = [F_1, I_{o2}, I_{o3}, I_{o4}], \tag{8}$$

D. DcUp Module

To further extract feature maps from section 3.3, this paper introduces an upsampling module called the DcUp module. Firstly, the input data is processed using a 1*1 convolutional kernel to reduce its dimension. Then, deformable convolution is introduced, replacing the sliding window sampling of traditional convolution with deformable sampling, adapting to various target deformations and scale changes. During deformable convolution, learnable offsets are incorporated, allowing each convolutional kernel to have variable sampling positions, enhancing the model's perception of targets. The reference formula is formula 9.

$$y(q_0) = \sum_{q_n \in R} w(q_n) \cdot x(q_0 + q_n + \Delta q_n) \tag{9}$$

where, q_0 represents a specific position in the feature map, a represents a 3x3 convolution kernel with positional offsets, $R = \{(-1,-1),(-1,0),(-1,1),...,(1,0),(1,1)\}$, Δq_n represents the offset values and w is the weight.

Finally, the receptive field of the feature map is expanded further, enhancing the network's understanding of input data and improving its performance in handling complex image objects. Transpose convolution is employed to upsample the feature map, resulting in an output image size of z, as specified in formula 10.

$$z = s \cdot (y_{in} - 1) - 2p + k \tag{10}$$

where, $s = 2$ represents the convolution kernel size stride, y_{in} is the size of the input feature map, $p = 1$ represents stri -de, and $k = 4$ is the size of the convolution kernel.

E. Loss function

This section introduces the loss functions used in the paper. The loss functions comprise three parts: the center point loss function, the offset loss function, and the width-height loss function.

The center point loss is utilized to assess the precision of center point predictions, and detailed references are available in formula 11.

$$\ell_{key} = -\frac{1}{N}\sum_{xyc} \begin{cases} (1-a_{xyc}^{pre})^{\alpha}\log(a_{xyc}^{pre}) & a_{xyc}^{gt} = 1 \\ (1-a_{xyc}^{gt})^{\beta}(a_{xyc}^{pre})^{\alpha}\log(1-a_{xyc}^{pre}) & otherwise \end{cases} \tag{11}$$

where, a_{xyc}^{gt} a represents the weight of the distance between the target center point and the x, y position, N is the number of keypoints, and c represents the number of classes. $a_{xyc}^{pre} \in [0,1]^{\frac{h}{R}\times\frac{w}{R}\times c}$ represents the confidence that the output pixel of the network is a center point, where R denotes the scaling factor. In this paper, we set the hyperparameters $\alpha = 2$ and $\beta = 4$.

The offset loss is employed to assess the accuracy of predicting the offset of the center point position. During the scaling process, the model generates the center point position's offset, thereby introducing the offset loss as referenced in formula 12.

$$\ell_{Off} = \frac{1}{N}\sum_{m}\left|\hat{k}_{\tilde{m}} - (\frac{m}{R} - \tilde{m})\right| \tag{12}$$

where, $\hat{k}_{\tilde{m}}$ represents the predicted offset of the model and m represents the center point. $\frac{m}{R} - \tilde{m}$ is the offset of the true center point calculated after downsampling, where $R = 4$ is the downsampling factor.

The width-height loss penalizes the model's inaccuracies in predicting the differences between the predicted and true width and height. This loss function facilitates a more accurate fitting of the target's size and shape by the model, as described in formula 13. where, C represents the number of classes,

$$\ell_{size} = \frac{1}{C}\sum_{i=1}^{C}\left|S_i^{pre} - S_i^{gt}\right| \tag{13}$$

$S_i^{gt} = \left(x_2^i - x_1^i, y_2^i - y_1^i\right)$ The width and height of a specific sample's ground truth. $S_i^{pre} \in \mathbb{R}^{\frac{W}{R}\cdot\frac{H}{R}\cdot 2}$ is the result of the network's regression. W is represents width, H represents height, R represents stride.

The total loss function of the model shown in formula 14, where $\lambda_{off}=1$ and $\eta_{size}=0.1$ represents the hyperparameter.

$$\ell_{\text{loss}}=\ell_{Key}+\lambda_{off}\ell_{off}+\eta_{size}\ell_{size}, \tag{14}$$

IV. Experimental Analysis

A. DataSets

VisDrone [17] is a comprehensive visual benchmark dataset designed for object detection. The dataset comprises 10 classes: pedestrian, person, bicycle, car, truck, van, tricycle, awning-tricycle, bus, and motor, with professional annotations. The dataset contains 6,471 training images, 548 validation images, and 3,190 testing images. Among them, 1,610 images are publicly available for testing, while 1,580 images are unopened and considered challenging. Images in the dataset have a maximum resolution of 2000×1500.

The VisDrone2019 dataset includes diverse aerial scenes, such as urban, traffic, rural, and coastal scenarios, with numerous interferences and complex backgrounds. The network model must accurately detect objects amidst these intricate backgrounds. Figure 3 shows that the dataset exhibits diverse object sizes, indicating a large number of small and medium-sized objects in the VisDrone2019 dataset. However, accurately detecting these small and medium-sized objects can be challenging, thus requiring the model to be sensitive to them.

In this paper, we define the sizes of the objects based on the MSCOCO dataset, where small objects have sizes less than 32×32, medium-sized objects have areas between 32×32 and 96×96, and large objects have sizes greater than 96×96.

B. Evaluation Metrics

We selected AP50, AP70, AP75, mAP[0.5:0.95], AP_{large}, AP_{medium}, and AP_{small} as evaluation metrics to assess the overall performance of the network.

Precision refers to the ratio of true positive samples to all samples classified as positive samples. The formula for calculating precision is given by formula 15.

$$Precision=\frac{TP}{TP+FP} \tag{15}$$

where, *TP* represents true positives, indicating the number of samples correctly predicted as positive by the model, while *FP* represents false positives, indicating the number of negative samples incorrectly predicted as positive by the model. Recall refers to the proportion of true positive samples out of all positive samples.

Recall is the proportion of correctly predicted positive samples among all positive samples. The calculation method for Recall can be found in formula 16.

$$Recall=\frac{TP}{TP+FN} \tag{16}$$

where, *FN* represents the positive samples incorrectly predicted as negative by the model.

Iou is used to quantify the overlap between the predicted box and the ground truth box. It is calculated by taking the intersection area of the predicted box and the ground truth box and dividing it by their union area, as show formula 17.

$$IOU=\frac{area_{pre}\cap area_{gt}}{area_{pre}\cup area_{gt}} \tag{17}$$

where, $area_{pre}$ represents the predicted box and $area_{gt}$ is ground truth.

Average precision (*AP*) is a crucial performance metric in object detection tasks. It ranges from $AP\in[0,1]$, with higher values indicating better model performance. In object detection tasks, various Iou thresholds are commonly used to calculate AP. Common thresholds include 0.5 and 0.75. AP_{50} and AP_{75} are calculated with all class-specific Iou thresholds set to 0.5 and 0.75, respectively. The calculation method for AP can be found in formula 18.

$$AP=\sum_{n=1}^{N}(R_n-R_{n-1})\cdot P_n \tag{18}$$

where, N represents the number of positive samples, R_n represents the recall of the top n samples, and P_n represents the precision of the top n samples.

The mean Average Precision (mAP) is the average of the AP values for all categories, and its formula is shown in formula 19.

$$mAP=\frac{1}{n}\sum_{i=1}^{n}AP_i, \tag{19}$$

C. Experiment Details

DUCAF-Net was developed using Pytorch 1.11. During the training phase, the model underwent 300 epochs on a 3080TI image processor. The initial learning rate was set at 6.25e−4, with a 10% reduction at epochs 150 and 200. We used a pretrained ResNet50, trained on the ImageNet dataset, as the backbone network. Data augmentation techniques, such as image flipping, scaling, and saturation adjustment, were applied to the input data. These techniques are proven effective for object detection. The post-processing shape of the model's input images is $3\times512\times512$.

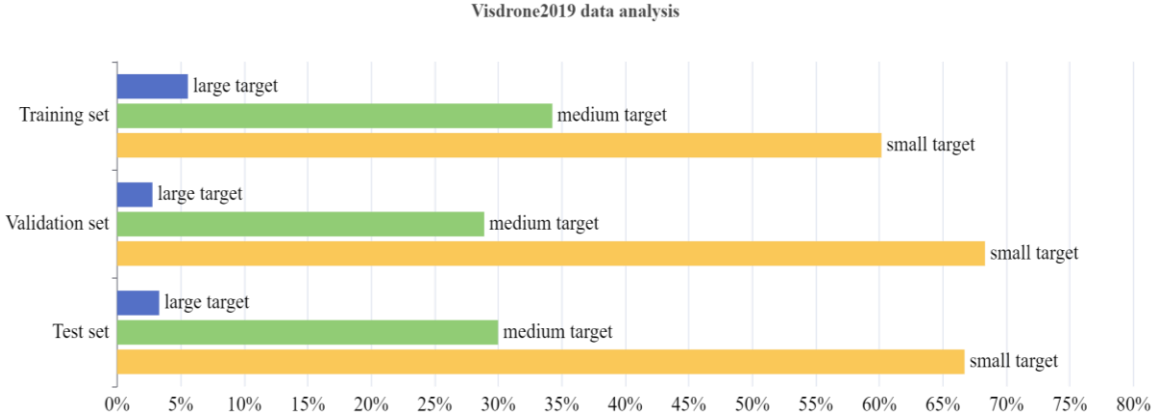


Figure 3 The data distribution graph of the VisDrone2019 dataset

D. Baseline

This paper compares a total of 11 baseline models, comprising single-stage, two-stage, anchor-based, and anchor-free object detection methods. Detailed information about these models is provided in the following sections.

1) Libra R-CNN[18] employs a novel region box regression method. In particular, it incorporates a learnable scale factor to adjust the size of region boxes and a learnable translation factor to modify their position, enabling adaptive resizing and repositioning of the region boxes to accommodate objects of different scales.

2) Yolov3 utilizes DarkNet-53 as the backbone network for feature extraction. By incorporating multiple detection heads at different scales, Yolov3 enhances detection accuracy and coverage. This approach enables Yolov3 to detect objects across various scales effectively.

3) In the first stage, Cascade-RCNN[19] uses a region proposal network to generate a large number of candidate regions. In the second stage, Cascade-RCNN employs cascaded detectors to refine and classify the candidate regions. Each cascaded detector comprises multiple sub-networks, and each sub-network further refines and adjusts the output of the previous sub-network to achieve more accurate detection results.

4) CenterNet[20] adopts DLA-34 as the backbone network. The DLA network features a multi-branch structure and inter-layer connections, demonstrating excellent performance in image feature extraction. CenterNet utilizes an efficient algorithm, unlike traditional anchor-based methods, to represent the object's position using a single center point. By predicting the coordinates, width, and height of this center point, it detects the object.

5) FCOS [21] employs a feature pyramid network to improve the algorithm's detection capabilities for objects of different sizes, as well as its ability to process high-resolution images. The key idea behind the FCOS network is to view object detection as a dense prediction task.

6) The ATSS [22] network utilizes a highly flexible strategy that employs statistical methods to assign positive and negative samples. By dynamically adjusting the sample selection strategy based on their importance, the network can better handle challenging and noisy samples, thereby improving the detection of hard-to-identify objects.

7) The TridentNet[23] network architecture comprises three branches, each having different convolution kernel sizes, to handle distinct receptive fields. However, these three branches share a common feature map. This allows each branch to concentrate on processing objects of varying scales, thus enhancing detection accuracy.

8) YOLOv5 utilizes an innovative adaptive convolution module that dynamically adjusts the size and shape of the convolutional kernel according to the targets' size and shape in the feature map, thus enhancing detection accuracy.

9) SAMFR-Cascade RCNN [24] employs an adaptive multi-scale fusion mechanism that selectively chooses multiple scales from the feature map and cascades them, enhancing the model's detection capability.

10)EfficientDet [25]utilizes a novel composite coefficient to optimize the network's complexity. This approach not only enhances the model's detection performance but also improves its accuracy.

11) FE-YOLOv5 [26] is composed of two parts: the first part is a cross-layer interaction module designed for shallow features, and the second part employs a cross-layer recombination approach to construct modules for deep features. This architecture enhances the detection capability for small objects

E. Comparative Experiments

We compares the methods from recent years. Please refer to Table 1 for details. The experiments show that the DUCAF-Net model demonstrates good performance compared to mainstream models in recent years.

DUCAF-Net demonstrates excellent overall performance compared to other popular one-stage object detection networks. For instance, compared to the original Yolov3, DUCAF-Net achieves a 66% improvement in detection accuracy, and compared to the original Yolov5, it achieves a 25.8% improvement in detection accuracy. DUCAF-Net achieves a 35.5% improvement in detection accuracy compared to CenterNet, a one-stage object detector with the DLA-34 backbone. DUCAF-Net achieves a 28% improvement in detection accuracy compared to the FCOS network. DUCAF-Net has achieved a 7.0% improvement in detection accuracy compared to the EfficientDet network. These results demonstrate that DUCAF-Net is a highly performant one-stage object detection network.

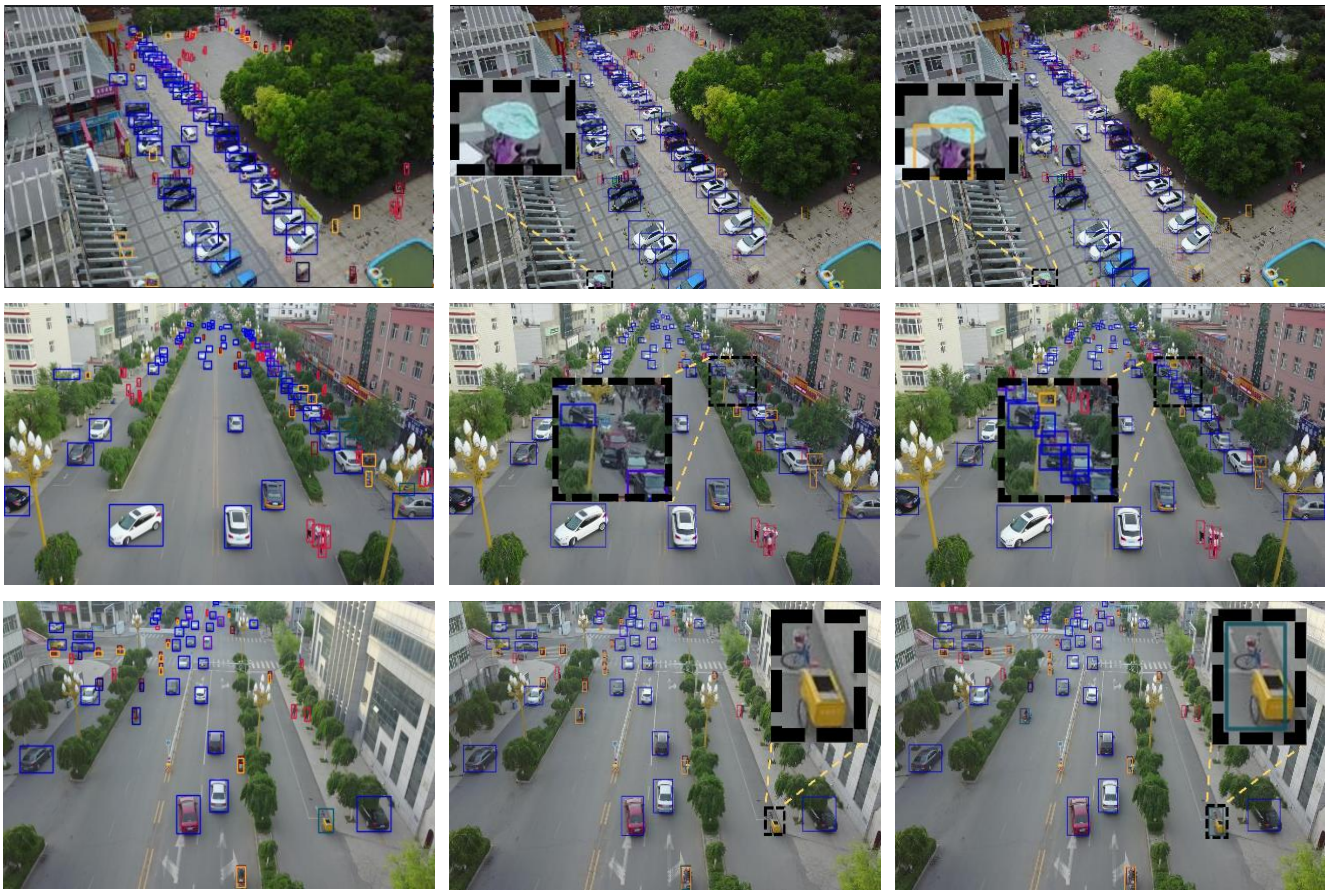
DUCAF-Net exhibits remarkable performance in detection accuracy compared to two-stage object detection networks. Compared to the original Cascade-RCNN, DUCAF-Net achieves a 42.4% improvement in detection accuracy. Compared to SAMFR-Cascade RCNN, DUCAF-Net achieves a 7.0% improvement in detection accuracy. Moreover, DUCAF-Net achieves a 15.7% improvement in detection accuracy compared to TridentNet. While DUCAF-Net does not surpass SAMFR-Cascade RCNN in AP50 scores, two-stage object detection algorithms have been proven to be more advantageous than one-stage algorithms in most cases, mAP score indicates DUCAF-Net performs better with a 13.5% improvement in detection accuracy.

Compared to feature-enhancement methods ATSS, DUCAF-Net demonstrates superior performance, achieving a 12.3% increase in mAP score. DUCAF-Net shows a 9.0% improvement in mAP score compared to FE-Yolov5. Nonetheless, FE-Yolov5 slightly outperforms DUCAF-Net in terms of AP_{large} score. FE-Yolov5's ability to use the GAU module contributes to its performance in detecting large objects. The module processes feature information from neighboring regions around each pixel, establishing long-range feature dependencies. However, despite this advantage, DUCAF-Net outperforms FE-YOLOv5 when comparing other evaluation metrics.

Overall, DUCAF-Net performs well. It exhibits excellent performance, especially in terms of APmedium score, with an additional improvement in APsmall score. These results indicate that DUCAF-Net performs well in detecting small and medium-sized objects. Moreover, the overall comprehensive performance of the network has been improved, further confirming the good performance of DUCAF-Net. To vividly demonstrate the effectiveness of the proposed method, we present a visual comparison of the results in Figure 4. The leftmost three images in Figure 4 display the annotation results of the original dataset. The middle three images represent the detection results of the CenterNet network. The rightmost three images show the detection results of the DUCAF-Net model. By comparing the original annotated data with the detection results of Cent

TABLE 1.
CONTRAST EXPERIMENT
Shows the performance comparison on the visdrone2019 test-det dataset, with bold indicating the best performance.

Method	mAP	AP_{50}	AP_{75}	AP_{small}	AP_{medium}	AP_{large}
Libra R-CNN[26]	14.70	24.60	15.40	6.10	25.20	31.40
Cascade-RCNN[27]	16.09	31.91	15.01	-	-	-
TridentNet[26]	19.80	35.00	19.50	11.40	29.60	36.60
SAMFR-Cascade RCNN[28]	20.18	**40.03**	18.42	-	-	-
EfficientDet (B5)[28]	21.40	38.60	20.20	-	-	-
Yolov3[27]	13.80	30.43	11.18	-	-	-
CenterNet[27]	16.90	32.10	15.50	-	-	-
FCOS[26]	17.90	30.40	18.30	9.20	27.60	35.40
Yolov5[26]	18.20	32.90	17.40	10.40	27.00	35.30
ATSS[26]	20.40	33.80	20.90	11.60	31.70	36.70
FE-YOLOv5[26]	21.00	37.00	20.70	13.20	29.50	**39.10**
DUCAF-Net(ours)	**22.91**	39.38	**23.10**	**13.52**	**33.30**	38.92



ORIGINAL ANNOTATION CENTERNET DUCAF-NET

Figure 4 Comparison of experimental visualization results. Displaying the original annotations of different roads in aerial images and the detection results of two different models. The black boxes indicate the detection differences between the two models, while the different color boxes within the black boxes represent the detected categories.

TABLE 2.
THE ABLATION EXPERIMENTS OF DUCAF-NET
The checkmark (√) indicates that a module was added and the hyphen (-) indicates that a module was not added.

Resnet50	MCAF	CA	DcUp	mAP	AP50	AP75	AP_{small}	AP_{medium}	AP_{large}
√	-	-	-	18.4	34.5	18.0	9.3	29.2	38.1
√	√	-	-	19.1	35.0	18.6	9.8	29.3	37.9
√	-	√	-	19.6	35.4	19.2	9.5	30.4	40.1
√	-	-	√	18.9	33.9	18.7	9.7	29.5	36.1
√	√	√	-	21.8	38.0	22.4	12.9	32.2	36.0
√	√	-	√	22.3	38.8	23.0	13.0	32.8	38.9
√	-	√	√	20.0	36.1	19.2	10.2	29.8	37.5
√	√	√	√	22.9	39.4	23.1	13.5	33.3	38.8

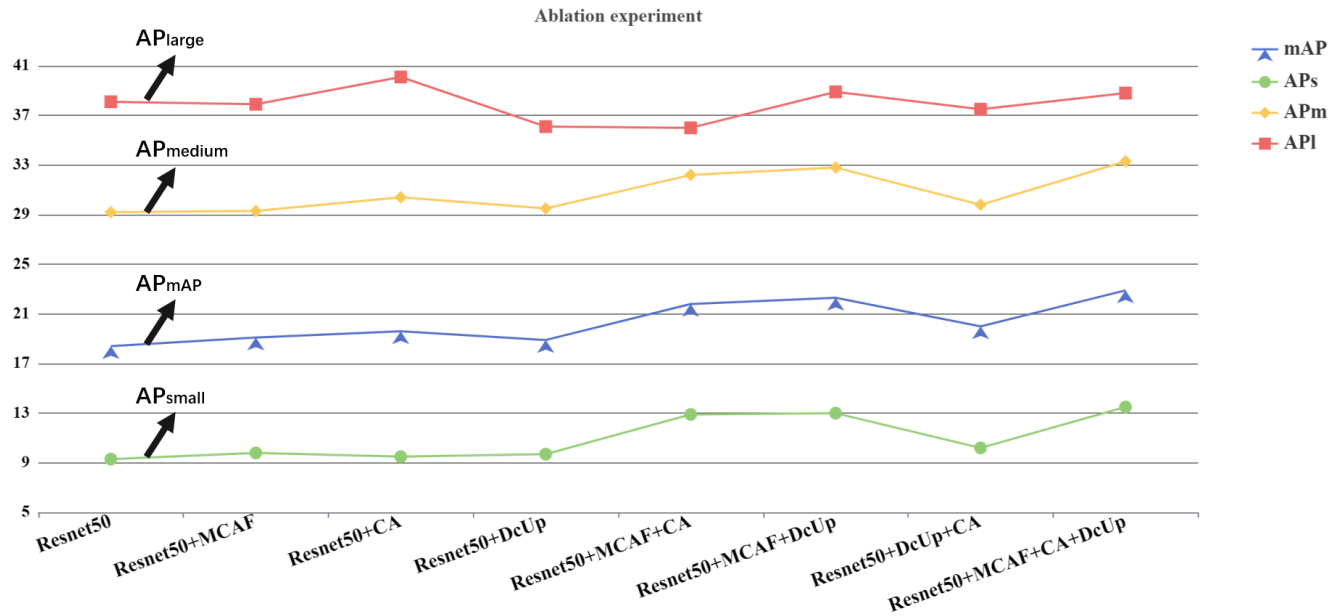


Figure 5 Ablation Experiment Result Chart. ResNet50+MCAF +CA+DcUp achieved the best performance.

-erNet and DUCAF-Net,we observe that in the first row of images, DUCAF-Net detects more objects, such as the bicycle parked in the lower-left corner. In the second row of images, DUCAF-Net detects a higher number of densely-packed vehicle targets compared to the CenterNet detection results. From the comparison of the third row of images, we conclude that DUCAF-Net successfully detects tricycles, and its detection results are closer to the original annotations. Based on the comprehensive analysis above, we draw the conclusion that the proposed DUCAF-Net exhibits significant superior performance in detecting small and medium-sized objects.

F. Sensitivity Analysis

In this section, we conduct ablation experiments on DUCAF-Net to verify its effectiveness. We perform parameter tuning experiments on the three main components of DUCAF-Net, namely the CA module, MCAF module, and DcUp module. Firstly, we add different modules to the backbone network in various combinations to verify their performance. Finally, we add all the modules to the backbone network to verify their performance. The results of the ablation experiments are shown in Table 2.

In cases of significant variation in object sizes, traditional attention mechanisms tend to overly focus on the channel dimensions of features. In contrast, the Coordinate Attention (CA) module performs attention computation in the spatial dimensions. By modeling and calculating attention based on the width and height directions of the objects, the CA module can more accurately focus on the positional information of the targets, leading to a better understanding of the image. The comparison results between the third row and the first row in Table 3 show that the introduction of the CA module has improved the mAP score by 1.2. This indicates that the incorporation of this module effectively enhances the detection performance.

The MCAF module (without CA module) enhances object detection capability and effectively mitigates the decrease in the overall model's detection ability caused by scene scale variations. While the MCAF module's introduction leads to a decrease in the network's ability to detect larger objects, it improves the overall performance. By comparing the data in the second row and the first row of Table 3, the mAP score increased by 0.7.

Comparing the fourth row with the first row in Table 3, we can observe that the introduction of the DcUp module improves the model's adaptability to variations in different scale targets. The DcUp module enables more accurate modeling of features from different regions, thereby enhancing the accuracy of feature expression. Despite a slight decrease in the model's detection of larger objects due to the introduction of the DcUp module, the overall performance is improved, resulting in an mAP score increase of 0.5.

Ablation experiments for MCAF+CA attention, MCAF+DcUp, and CA+DcUp were performed in rows 5, 6, and 7 of Table 3. The experimental results demonstrate that incorporating these modules has enhanced the overall model performance, outperforming the individual addition of each module. In particular, when compared to the baseline model, the MCAF+CA Attention network achieved a 3.4 increase in mAP score, the MCAF+DcUp network achieved a 4.1 increase in mAP score, and the CA+DcUp network achieved a 1.6 increase in mAP score.

The experimental results indicate that the performance of the eighth row (Resnet50+MCAF+DcUp+CA) is superior to rows one to seven, achieving an mAP of 22.9. The mAP score has improved by 4.5 compared to the baseline. From the analysis of the ablation experiment in Figure 5, it can be observed that incorporating all modules into the network yields the best performance.

Based on a comprehensive analysis, we have concluded that each module of DUCAF-Net performs well in the object detection task, and they demonstrate mutual dependence and interaction.

V. CONCLUSION

This paper proposes a method called DUCAF-Net to improve the detection capability of small objects in aerial photography datasets. The method uses the MCAF module to fuse feature maps with coordinate attention from various resolutions and dimensions. This enables learning spatial feature information with diverse dimensions and resolutions, thereby enhancing sensitivity to the positional information

of the feature maps. The DcUp module is employed to enhance the overall fusion effect by adjusting feature response weights for different coordinate positions, reducing feature diffusion and coupling in dense small object regions. Experimental results on the VisDrone dataset show the method performs well. In the future, we will keep exploring aerial photography datasets and optimizing object detection algorithms to enhance their application value in aerial and earth science fields.

REFERENCES

[1] X. Xu, "Research on a Small Target Object Detection Algorithm for Electric Transmission Lines Based on Convolutional Neural Network," *IAENG International Journal of Computer Science*, vol. 50, no. 2, pp. 375-380, 2023.

[2] H. Zhang and J. Zhao, "Traffic Sign Detection and Recognition Based on Deep Learning," *Engineering Letters*, vol. 30, no. 2, pp. 666-673, 2022.

[3] Z. Li *et al.*, "Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey," *Remote Sensing*, vol. 14, no. 10, p. 2385, 2022.

[4] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: A review," *Image and Vision Computing*, vol. 97, p. 103910, 2020/05/01/ 2020.

[5] V. Sharma and R. N. Mir, "A comprehensive and systematic look up into deep learning based object detection techniques: A review," *Computer Science Review*, vol. 38, p. 100301, 2020/11/01/ 2020.

[6] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837-128868, 2019.

[7] T. Liu, B. Pang, S. Ai, and X. Sun, "Study on visual detection algorithm of sea surface targets based on improved yolov3," *Sensors (Switzerland)*, vol. 20, no. 24, pp. 1-14, 2020.

[8] S. H. Kang and J. S. Park, "Aligned Matching: Improving Small Object Detection in SSD," *SENSORS*, vol. 23, no. 5, p. 2589, MAR 2023, Art. no. 2589.

[9] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery With Improved YOLOv5 Based on Transfer Learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8085-8094, 2022.

[10] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, October 27, 2019 - November 2, 2019*, Seoul, Korea, Republic of, 2019, vol. 2019-October, pp. 8231-8240: Institute of Electrical and Electronics Engineers Inc.

[11] Z. Sun *et al.*, "An Anchor-Free Detection Method for Ship Targets in High-Resolution SAR Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7799-7816, 2021.

[12] A. Jawaharlalnehru *et al.*, "Target Object Detection from Unmanned Aerial Vehicle (UAV) Images Based on Improved YOLO Algorithm," *ELECTRONICS*, vol. 11, no. 15, p. 2343, AUG 2022, Art. no. 2343.

[13] X. Luo, Y. Wu, and L. Zhao, "YOLOD: A Target Detection Method for UAV Aerial Imagery," *Remote Sensing*, vol. 14, no. 14, p. 3240, 2022.

[14] H. Fang, M. Xia, G. Zhou, Y. Chang, and L. Yan, "Infrared Small UAV Target Detection Based on Residual Image Prediction via Global and Local Dilated Residual Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[15] Y. Guo, X. Tong, X. Xu, S. Liu, Y. Feng, and H. Xie, "An Anchor-Free Network With Density Map and Attention Mechanism for Multiscale Object Detection in Aerial Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021, June 19, 2021 - June 25, 2021*, Virtual, Online, United states, 2021, pp. 13708-13717: IEEE Computer Society.

[17] D. Du *et al.*, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *17th IEEE/CVF International Conference on Computer Vision Workshop, ICCVW 2019, October 27, 2019 - October 28, 2019*, Seoul, Korea, Republic of, 2019, pp. 213-226: Institute of Electrical and Electronics Engineers Inc.

[18] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16, 2019 - June 20, 2019*, Long Beach, CA, United states, 2019, vol. 2019-June, pp. 821-830: IEEE Computer Society.

[19] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in *31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18, 2018 - June 22, 2018*, Salt Lake City, UT, United states, 2018, pp. 6154-6162: IEEE Computer Society.

[20] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, October 27, 2019 - November 2, 2019*, Seoul, Korea, Republic of, 2019, vol. 2019-October, pp. 6568-6577: Institute of Electrical and Electronics Engineers Inc.

[21] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, October 27, 2019 - November 2, 2019*, Seoul, Korea, Republic of, 2019, vol. 2019-October, pp. 9626-9635: Institute of Electrical and Electronics Engineers Inc.

[22] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, June 14, 2020 - June 19, 2020*, Virtual, Online, United states, 2020, pp. 9756-9765: IEEE Computer Society.

[23] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, October 27, 2019 - November 2, 2019*, Seoul, Korea, Republic of, 2019, vol. 2019-October, pp. 6053-6062: Institute of Electrical and Electronics Engineers Inc.

[24] H. Wang *et al.*, "Spatial attention for multi-scale feature refinement for object detection," in *17th IEEE/CVF International Conference on Computer Vision Workshop, ICCVW 2019, October 27, 2019 - October 28, 2019*, Seoul, Korea, Republic of, 2019, pp. 64-72: Institute of Electrical and Electronics Engineers Inc.

[25] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, June 14, 2020 - June 19, 2020*, Virtual, Online, United states, 2020, pp. 10778-10787: IEEE Computer Society.

[26] M. Wang *et al.*, "FE-YOLOv5: Feature enhancement network based on YOLOv5 for small object detection," *Journal of Visual Communication and Image Representation*, vol. 90, p. 103752, 2023.

[27] J. Liao, Y. Liu, Y. Piao, J. Su, G. Cai, and Y. Wu, "GLE-Net: A Global and Local Ensemble Network for Aerial Object Detection," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, p. 2, 2022.

[28] V. Chalavadi, P. Jeripothula, R. Datla, S. B. Ch, and K. M. C, "mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognition*, vol. 126, p. 108548, 2022.