Article

# DMF-YOLO: Dynamic Multi-Scale Feature Fusion Network-Driven Small Target Detection in UAV Aerial Images

Xiaojia Yan , Shiyan Sun , Huimin Zhu * , Qingping Hu , Wenjian Ying , Yinglei Li

Posted Date: 11 June 2025

doi: 10.20944/preprints202506.0837.v1

Keywords: DMF-YOLO; UAV aerial images; target detection; feature extraction; multi-scale objects

*Article*

# DMF-YOLO: Dynamic Multi-scale Feature Fusion Network-Driven Small Target Detection in UAV Aerial Images

**Xiaojia Yan, Shiyan Sun, Huimin Zhu *, Qingping Hu, Wenjian Ying and Yinglei Li**

Graduate school, Naval University of Engineering, 717 Jiefang Road, Qiaokou District, Wuhan 430030, China

* Correspondence: D23182604@nue.edu.cn

**Abstract:** Target detection in UAV aerial images has found increasingly widespread applications in emergency rescue, maritime monitoring, and environmental surveillance. However, traditional detection models suffer significant performance degradation due to challenges including substantial scale variations, high proportions of small targets, and dense occlusions in UAV-captured images. To address these issues, this paper proposes DMF-YOLO, a high-precision detection network based on YOLOv10 improvements. First, we design Dynamic Dilated Snake Convolution (DDSConv) to adaptively adjust the receptive field and dilation rate of convolution kernels, enhancing local feature extraction for small targets with weak textures. Second, we construct a Multi-scale Feature Aggregation Module (MFAM) that integrates dual-branch spatial attention mechanisms to achieve efficient cross-layer feature fusion, mitigating information conflicts between shallow details and deep semantics. Finally, we propose an Expanded Window-based Bounding Box Regression Loss Function (EW-BBRLF), which optimizes localization accuracy through dynamic auxiliary bounding boxes, effectively reducing missed detections of small targets. Experiments on the VisDrone2019 and HIT-UAV datasets demonstrate that DMF-YOLOv10 achieves 50.1% and 81.4% mAP50, respectively, significantly outperforming the baseline YOLOv10s by 27.1% and 2.6%, with parameter increases limited to 24.4% and 11.9%. The method exhibits superior robustness in dense scenarios, complex backgrounds, and long-range target detection. This approach provides an efficient solution for UAV real-time perception tasks and offers novel insights for multi-scale object detection algorithm design.

**Keywords:** DMF-YOLO; UAV aerial images; target detection; feature extraction; multi-scale objects

## 1. Introduction

With the rapid development of unmanned aerial vehicle (UAV) technology, its applications in emergency rescue [1], marine monitoring [2], environmental surveillance [3], and various other fields have become increasingly widespread. Particularly in the realm of information perception, UAVs leverage their advantages of flexible deployment, high mobility, and broad field of view to effortlessly acquire detailed information about target areas. However, dense multi-scale object detection remains one of the most challenging issues in object detection tasks due to factors such as small target sizes, large quantities, significant scale variations in UAV aerial imagery, resolution limitations, varying illumination conditions, and target occlusion [4–6]. This not only demands detection algorithms to exhibit high accuracy and robustness but also requires real-time processing and analysis of massive image data to meet the urgent need for rapid response in practical applications [7].

In the field of object detection, deep learning-based algorithms have become mainstream and are primarily divided into two-stage detection algorithms and single-stage detection algorithms based on their workflow. Two-stage detection algorithms [8], such as the R-CNN series [9], first generate region proposals and then perform feature extraction and classification on these regions. Although this approach achieves higher detection accuracy, it incurs higher computational costs and slower speeds, making it unsuitable for scenarios requiring rapid responses. The Faster R-CNN [10] algorithm significantly improves detection speed by introducing a Region Proposal Network, but it still requires classification and bounding box regression for each candidate region, which limits its performance in real-time applications. In contrast, single-stage detection algorithms [11], such as the YOLO (You Only Look Once) series [12] and SSD series [13], directly predict object classifications and bounding boxes on the entire image without generating region proposals, resulting in faster detection speeds suitable for real-time applications. Notably, the YOLO series has gained attention for its rapid detection speed and balanced performance, maintaining detection accuracy while ensuring speed. However, single-stage algorithms generally lag slightly behind two-stage algorithms in precision, particularly in small object detection.

Current mainstream object detection algorithms, primarily designed for natural scenes, often underperform when directly applied to UAV aerial imagery. This performance gap stems from significant differences between natural and aerial scenarios: drastic variations in object scales, high proportions of small objects, dense object distributions, and severe mutual occlusion. These factors collectively lead to a notable decline in detection accuracy [14]. To address these challenges, researchers have proposed various model improvement strategies. Lin et al. [15] pioneered the Feature Pyramid Network (FPN) architecture, enabling multi-scale feature interaction through cross-layer connections, which established a foundational framework for efficient feature fusion. Building on FPN, Liu et al. [16] optimized feature propagation paths by proposing the Path Aggregation Network (PANet), which significantly enhanced the utilization of low-level features through a bottom-up augmentation path, though its performance remains highly sensitive to image quality. Deng et al. [17] introduced a progressive scale transformation method combined with a global-local fusion mechanism, effectively boosting small object detection performance, but this approach shows clear limitations in tracking medium-to-large objects. Cai et al. [18] improved detection accuracy by integrating a coordinate attention mechanism into the YOLOv4-tiny model, yet its generalization capability remains insufficient for cross-scene tracking tasks. Zhu et al. [19] proposed a detection architecture combining multi-transformer prediction heads with CBAM attention, significantly optimizing small object detection, though its high computational complexity compromises real-time efficiency. For model efficiency optimization, Ma et al. [20] innovatively employed channel splitting and recombination techniques, enhancing cross-scale feature interaction while achieving adaptive multi-level feature fusion. Sandler et al. [21] designed an inverted residual structure with high-dimensional channel expansion, which improves feature representation while maintaining model lightweightness, but exhibits significant accuracy degradation in occluded scenarios. These studies demonstrate that balancing real-time performance with robust multi-scale object detection remains a critical technical bottleneck requiring breakthroughs in UAV image-based target detection.

To address the aforementioned challenges, this paper proposes an enhanced algorithm named DMF-YOLOv10 based on the YOLOv10s framework, specifically designed for UAV aerial image detection scenarios. The primary innovative contributions of this study are manifested in the following three aspects:

(1) Innovative Design of Dynamic Dilated Serpentine Convolution (DDSConv) Layer: This layer dynamically adjusts the dilation rate of convolutional kernels according to the scale variations of input features, enabling adaptive reshaping of receptive fields. This mechanism effectively captures local features of small targets in aerial images. The improvement specifically addresses the limitations of traditional convolutional layers in handling weak texture features and low pixel density in aerial imagery, thereby enhancing the extraction of discriminative deep feature representations.

(2) Multi-scale Feature Aggregation Module (MFAM): A dual-branch feature interaction strategy is proposed to achieve multi-scale information complementarity by fusing high-resolution detail features with deep semantic features. The module employs dual-dimensional spatial attention to dynamically weight fused features, effectively suppressing redundant information interference. Compared to traditional fusion methods, MFAM significantly reduces information loss while enhancing feature representation capability through the integration of spatial attention and adaptive

weight allocation mechanisms, thereby providing more discriminative multi-scale representations for detection heads.

(3) MFAM-Neck Network for Aerial Objects: To address the challenges of large size variations and clustered small targets in aerial images, a dedicated neck network based on MFAM is designed for feature fusion. This architecture adopts a phased fusion strategy that reconstructs fusion pathways by associating dual-scale features, enabling fine-grained feature enhancement. Combined with MFAM's spatial co-optimization capability, the model significantly improves localization accuracy for multi-scale targets, particularly micro-scale objects.

(4) Extended Window-based Bounding Box Regression Loss Function (EW-BBRLF): Inspired by the auxiliary bounding box acceleration mechanism in Inner-IoU, this loss function integrates the direction-aware advantages of Complete Intersection over Union (CIoU) loss with the scale sensitivity of Ln norm. By adaptively adjusting auxiliary bounding box dimensions through a dynamic scaling coefficient, it enhances localization precision and detection accuracy.

The paper is organized as follows: Section 2 reviews recent advancements in multi-scale object detection for aerial imagery. Section 3 presents the improved model proposed for small object detection in UAV images, detailing the model architecture and operational principles of related modules. Section 4 outlines the experimental environment and parameter configurations, followed by test results on VisDrone2019 and HIT-UAV datasets, including ablation studies, comparative evaluations, and visualization experiments designed to validate the effectiveness of the proposed method. Section 5 concludes the paper and discusses potential directions for future research.

## 2. Related Work

Feature extraction in UAV aerial imagery is critical for object detection, requiring solutions to challenges such as illumination variations and scale discrepancies. Feature fusion addresses small object disappearance and semantic insufficiency by integrating multi-scale information, thereby enhancing model adaptability to complex scenes. For sample imbalance and small object detection, well-designed loss functions optimize model training and improve detection accuracy. The synergistic integration of these three components significantly enhances UAV aerial image analysis performance, playing a vital role in optimizing model detection speed and precision.

### *2.1. Feature Extraction*

UAV aerial images pose unique challenges for effective feature extraction due to their special shooting perspectives, large target scale variations, and highly complex backgrounds:

- Low-resolution small targets are vulnerable to noise interference.
- Densely arranged targets suffer from feature adhesion.
- High-angle shooting induces significant geometric distortions.

Traditional convolutional networks with fixed receptive fields struggle to adapt to the dynamic characteristics of aerial scenarios [22], necessitating lightweight and adaptive feature extraction methods that balance small target sensitivity with computational efficiency.

To address these issues:

Dai et al. [23] developed deformable convolutional networks (DCN) using dynamically adjustable offset mechanisms, significantly enhancing convolutional kernel spatial adaptability. However, multi-layer deformable structures increase model parameters by approximately 23%, escalating computational resource consumption during training.

Du et al. [24] proposed a sparse convolution scheme that applies channel pruning optimization to detection heads, reducing computational complexity while maintaining detection accuracy. However, this approach compromises feature discriminability in complex backgrounds.

Wang et al. [25] constructed an elastic receptive field model using deformable convolution as a base operator, breaking traditional geometric constraints. Experiments demonstrated a 2.1×

expansion in effective receptive fields, yet failed to establish mechanisms linking receptive fields to small target spatial distributions.

Qi et al. [26] developed DSCNet with serpentine convolution structures, improving segmentation accuracy for vascular networks. However, its topological feature extraction mechanism is susceptible to irregular shape interference in generic object detection scenarios, and dynamic kernel adjustment stability requires optimization.

Niu et al. [27] integrated attention mechanisms into dynamic snake convolution architectures, mitigating insufficient positive sample feature capture in traditional methods and improving road crack detection completeness.

Chen et al. [28] proposed the MDSC-YOLOv9 model for steel surface defect detection. This method embeds Local-enhanced Positional Encoding (LePE) attention mechanisms and improved upsampling modules into feature fusion networks to enhance small target feature representation and sensitivity to narrow/elongated defects.

While dynamic snake convolution effectively extracts local features of elongated and scale-varying structures, it cannot adjust receptive field sizes, leading to offset imbalance in extreme scale variations or irregular target scenarios, thus failing to focus on small target local features.

### *2.2. Feature Fusion*

The multi-scale target distribution characteristics of aerial imagery demand feature fusion mechanisms with cross-layer semantic correlation capabilities. Traditional methods often suffer from small target feature loss due to information attenuation during the fusion of deep abstract features and shallow detail features [29].

To address these challenges:

Zhang et al. [30] proposed the DRF-SGA framework, which employs a self-guided attention mechanism for cross-modal feature calibration. This approach improves detection accuracy in low-light conditions, though its recall rate for targets smaller than 16×16 pixels requires further enhancement.

Yang et al. [31] introduced the QueryDet acceleration algorithm, utilizing a feature pyramid pre-screening mechanism to boost inference speed on the COCO dataset. However, its false detection rate exceeding thresholds triggers cascading effects of invalid feature processing.

Duan et al. [32] innovated the CenterNet model with a key-point triplet supervision strategy, achieving target localization through center-point heatmap regression. While effective for medium-scale object detection, its small target feature representation capability needs strengthening.

Wang et al. [33] recently developed the Gold-YOLO architecture, which integrates multi-scale feature collection-distribution modules and channel attention mechanisms to enhance feature reuse rates. This design maintains the integrity of original information across all feature scales while achieving an ideal balance between computational latency and accuracy.

### *2.3. Loss Functions*

Loss functions are pivotal for guiding model learning, yet traditional detection models often struggle to precisely model the complex spatial relationships between targets and backgrounds in UAV aerial images characterized by dense target distributions, diverse complex backgrounds, and large-scale target variations [34].

Recent advances and limitations include:

Zhu et al. [35] developed the FSAF framework, which integrates anchor-free detection architecture with feature selection modules to enable adaptive cross-layer feature fusion. However, its strong dependence on data distribution leads to degraded generalization in non-typical scenarios.

Chen et al. [36] proposed KD Loss for remote sensing object detection, enhancing feature discriminability by mapping features into high-dimensional Hilbert spaces. This approach, however, inadequately addresses the dense clustering characteristics of small targets.

Gevorgyan et al. [37] designed SIoU Loss, reconstructing the loss computation framework with directional alignment constraints. This metric risks failure when predicted and ground-truth boxes share aspect ratios but differ substantially in scale.

Zhang et al. [38] introduced ATSS (Adaptive Training Sample Selection), balancing anchor-based and anchor-free methods through dynamic sample screening. Its performance remains sensitive to hyperparameter tuning and training set statistics.

Ma et al. [39] created MPDIoU, resolving computational flaws in equal-proportion scale-discrepant cases. Nevertheless, it still struggles with small-scale target detection tasks.

Zhang et al. [40] proposed Inner-IoU, optimizing regression through multi-scale auxiliary bounding boxes. While improving convergence efficiency, it insufficiently models the impact of spatial distance on regression accuracy.

## 3. Proposed Model

### *3.1. Overview of YOLOv10*

As a state-of-the-art achievement in real-time object detection, YOLOv10 incorporates innovative improvements while inheriting the advantages of its predecessors. The algorithm not only optimizes detection capabilities but also extends to multi-task support including classification, segmentation, and tracking. Its outstanding performance and architectural adaptability have garnered significant attention in the computer vision community [41–43]. The YOLOv10s baseline model selected in this study employs a three-stage architecture: Backbone (feature extraction layer), Neck (feature fusion layer) and Head (prediction output layer). The network architecture is illustrated in Figure 1.

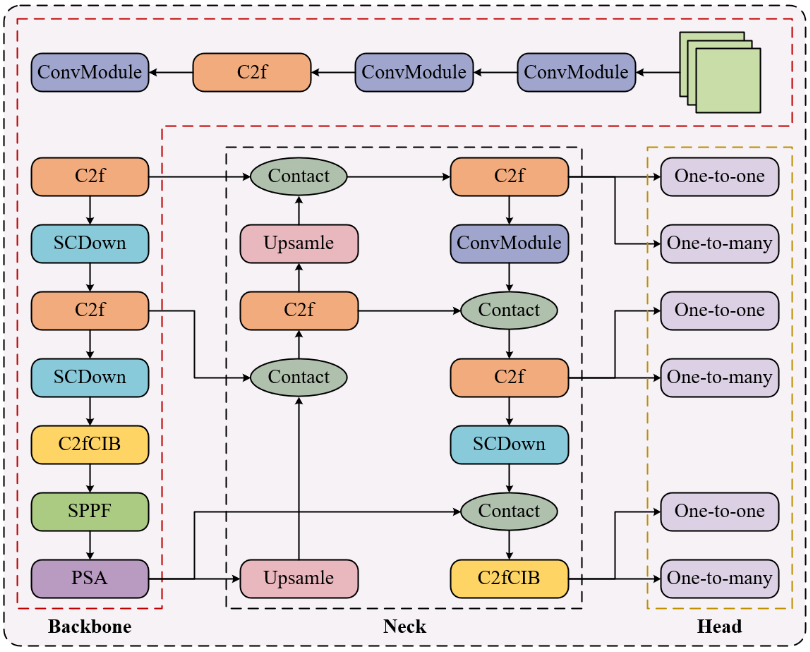


**Figure 1.** YOLOv10 network structure.

The backbone architecture builds upon Darknet-53 by integrating YOLOv8's CSPLayer_2Conv (C2f) module for residual connections, while innovatively employing a Spatial-Channel Decoupled Downsampling (SCDown) module to enhance downsampling efficiency. To address redundancy caused by repetitive modules in traditional architectures across stages, the model implements an efficiency-accuracy equilibrium design strategy: it introduces Compact Inverted Residual Blocks (CIB) to optimize fundamental units, and dynamically adjusts network depth through a rank analysis-guided adaptive module configuration scheme. For unified multi-scale feature representation, the model retains YOLOv8's Spatial Pyramid Pooling Fusion (SPPF) module and innovatively appends Position-Sensitive Attention (PSA) after SPPF, effectively enhancing feature expressiveness while maintaining low computational overhead.

The neck adopts an enhanced variant of the Bidirectional Feature Pyramid Network (BiFPN) for multi-scale feature fusion. The feature pyramid network combines deep semantic information with shallow detail features through a top-down propagation path, significantly improving multi-scale target representation via hierarchical feature integration. As a complementary optimization mechanism, a reverse feature fusion pathway establishes a bottom-up propagation channel to refine local detail integration.

The head employs a decoupled prediction structure that separates classification and regression tasks into two independent branches. Each branch contains prediction modules composed of 3×3 and 1×1 convolutions, with dynamic label assignment strategies introduced to optimize positive/negative sample allocation. The anchor-free design eliminates preset anchor boxes from traditional approaches. Convolution-based feature mapping layers directly output target geometry parameters (center coordinates, dimensions) and class probability distributions. This end-to-end prediction mechanism simplifies the detection pipeline while enhancing adaptability to target deformation and scale variations.

However, the YOLOv10 baseline model exhibits limitations in small target detection tasks for drone aerial imagery, with core challenges stemming from insufficient adaptation of feature extraction, feature fusion, and loss function design to aerial scene characteristics. During feature extraction, while YOLOv10 constructs high-level semantic features through deep convolutional networks with progressive downsampling, this process significantly compresses the spatial resolution of feature maps, causing effective pixel information of small targets to nearly vanish in high-level features. In drone imagery, densely distributed small targets (e.g., vehicles, pedestrians) typically occupy only tens of pixels. The large receptive fields of deep networks, though beneficial for large object detection, excessively dilute local detail features of small targets through repeated downsampling, leading to irreversible loss of critical texture and shape information.

Regarding feature fusion mechanisms, although the model achieves multi-scale feature interaction through feature pyramid networks (FPN), it fails to effectively reconcile the conflict between low-semantic shallow features and low-resolution deep features. Aerial target detection requires simultaneous reliance on high-resolution features for precise localization and deep features for contextual reasoning. Existing fusion strategies may inadequately extract multi-scale contextual relationships for small targets due to insufficient response from channel attention mechanisms to low signal-to-noise small target features, or information attenuation during cross-scale feature concatenation.

Furthermore, the loss function design shows bias in small target optimization. Intersection over Union (IoU)-based localization loss demonstrates insufficient sensitivity to coordinate variations in tiny bounding boxes, with gradient updates prone to local optima when target sizes significantly deviate from anchor priors. The dynamic balance mechanism between classification and localization losses also lacks adaptive adjustment for the extreme positive-negative sample imbalance inherent to small targets, causing the model to neglect low-confidence small target predictions. These factors collectively constrain the model's capacity for comprehensive small feature mining and precise regression in aerial scenarios.

### *3.2. Proposed Method*

#### 3.2.1. Overall Network Architecture

Aerial images often contain numerous small targets that occupy limited proportions and pixel compositions in the image, typically exhibiting slender columnar geometric shapes. Current methods primarily rely on convolutional and pooling layers to extract high-level feature information related to targets [44]. However, as convolutional layers are progressively stacked, feature map dimensions continuously shrink and resolutions degrade, causing small target information to be easily overlooked[45].

To address the limitations of YOLOv10 in detecting multi-scale targets, this paper proposes an innovative model, DMF-YOLO, designed to enhance the efficiency and accuracy of small target detection. The model improves adaptive multi-scale feature extraction capabilities through several modules. In the backbone, a Dynamic Dilated Snake Convolution (DDSConv) is designed to dynamically adjust the receptive field shape of convolutional kernels, enabling adaptive extraction of local features for multi-scale targets. This capability allows precise capture of critical information, effectively addressing feature extraction challenges in aerial images under conditions of weak local structures, sparse pixels, and complex background interference. For the neck design, a Multi-scale Feature Aggregation Module (MFAM) is constructed to integrate multi-layer feature maps, balancing robust semantic information with rich detail. This dual-branch architecture effectively preserves both local details and global contextual modeling. By introducing a spatial attention mechanism, the module strengthens the kernel's adaptive capacity to detect target-related features across spatial regions, thereby accommodating features of varying sizes and shapes across samples. Building on this, the MFAM-Neck further enhances multi-scale feature fusion, improving the network's perception of objects at different scales. The detection head retains the original YOLOv10 design. The refined network architecture is illustrated in Figure 2.

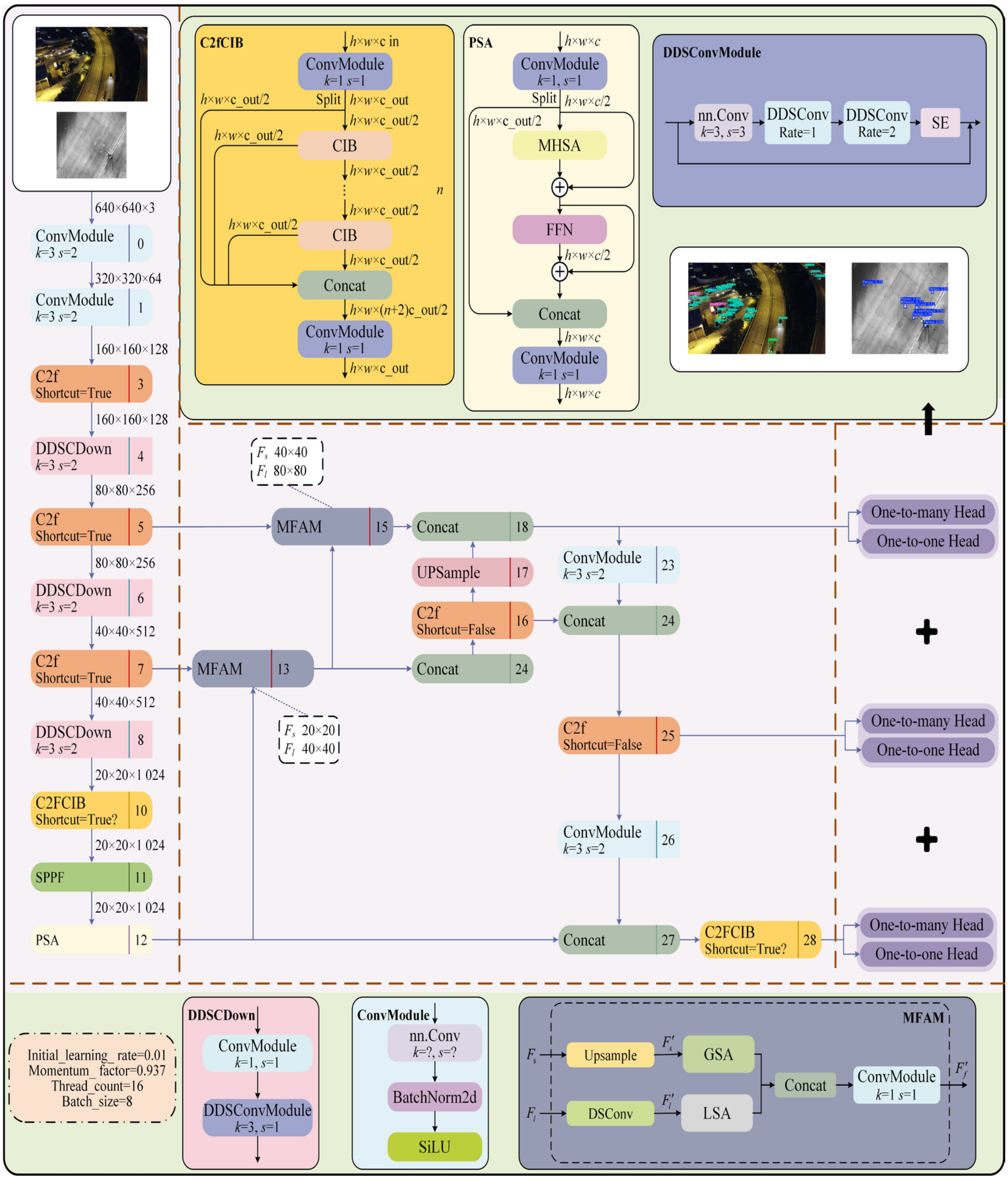


**Figure 2.** DMF-YOLO network structure.

### 3.2.2. Dynamic Dilated Snake Convolution (DDSConv)

Dynamic Snake Convolution effectively enhances perception of geometric structures by adaptively focusing on local features of thin and curved tubular shapes, achieving superior segmentation results in cardiac vessel datasets and the Massachusetts road dataset [26]. Similar to blood vessels and road networks, drone-observed targets also exhibit slender and highly variable characteristics, with lengths and widths showing complex variations in images [46]. However, Dynamic Snake Convolution cannot adaptively adjust the receptive field size, making it prone to offset imbalance under extreme scale variations or irregular aerial targets, thereby failing to focus on local features of small targets.To address these issues, we propose the Dynamic Dilated Snake Convolution (DDSConv), which dynamically adjusts the receptive field shape of convolutional kernels based on different dilation rates. Additionally, it adaptively focuses on local features of minute targets to more precisely capture their critical information. This design effectively resolves challenges related to fragile local structures, limited pixel counts, and interference from complex ground background information.

To enable convolutional kernels to more flexibly focus on complex geometric features of targets, DDSConv introduces deformation offsets $\Delta$. However, when the model is allowed to freely select deformation offsets, the receptive field tends to deviate from targets, particularly when processing slender structures. To address this, an iterative process with continuity constraints is implemented to prevent excessive divergence in detection results. During each convolutional operation, the model uses previous positions as references to sequentially select the next observation points, ensuring that attention balances flexibility and continuity for each target being processed. This approach facilitates the extraction of richer critical features and enhances model performance.

DDSConv samples input feature maps in the form of continuous stochastic grids, with the mathematical principles detailed as follows: Given a standard 2D convolution coordinate system $\mathbf{K}$ where the central coordinates are defined as $\mathbf{K}_i = (x_i, y_i)$, and other grid positions are denoted as $\mathbf{K}_{i\pm c} = (x_{i\pm c}, y_{i\pm c})$ , $c \in \{1, 2, \cdots, n\}$. The selection of each grid position $\mathbf{K}_{i\pm c}$ in the convolution kernel follows a recursive process. Starting from the central position $\mathbf{K}_i$, the position of distant grids $\mathbf{K}_{i\pm c}$ depends on the prior grid $\mathbf{K}_{i\pm c-1}$. The position of each grid is incrementally adjusted relative to its predecessor by adding an offset $\Delta = \{\delta \mid \delta \in [-1,\ 1]\}$. Given a central distance $c$, the positional variation of the grid is defined as $c\delta$. Different types of convolution operations are illustrated in Figure 3.

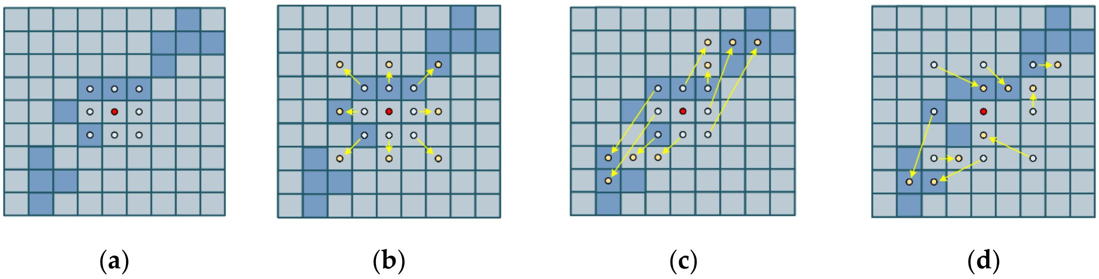


**Figure 3.** Different types of convolution operations. (**a**) Conv; (**b**) DConv; (**c**) DSConv; (**d**) DDSConv.

Based on the principles above, the variation of DDSConv in the axis direction is illustrated in Figure 4(a), with the calculation formula expressed as

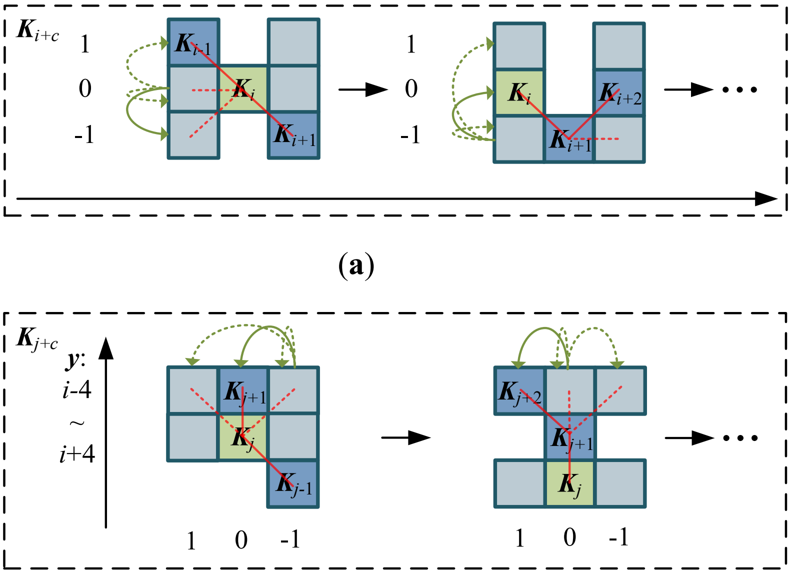
$$\boldsymbol{K}_{i\pm c} = \begin{cases} \left(x_{i+c}, y_{i+c}\right) = \left(x_i + c + \dfrac{n-1}{2} \times d, y_i + \sum\limits_{i}^{i+c} \Delta y + \dfrac{n-1}{2} \times d\right) \\ \left(x_{i-c}, y_{i-c}\right) = \left(x_i - c - \dfrac{n-1}{2} \times d, y_i + \sum\limits_{i-c}^{i} \Delta y + \dfrac{n-1}{2} \times d\right) \end{cases} \quad (1)$$

where $\sum\limits_{i}^{i+c}$ denotes the summation along the positive semi-axis of the *x*-axis, $\sum\limits_{i-c}^{i}$ represents the summation along the negative semi-axis of the *x*-axis, $n$ is the kernel size, and $d$ is the dilation factor.

The variation in the *y*-axis direction is illustrated in Figure 4(b), with the calculation formula expressed as

$$\boldsymbol{K}_{j\pm c} = \begin{cases} \left(x_{j+c}, y_{j+c}\right) = \left(x_j + \sum\limits_{j}^{j+c} \Delta x + \dfrac{n-1}{2} \times d, y_j + c\right) \\ \left(x_{j-c}, y_{j-c}\right) = \left(x_j + \sum\limits_{j-c}^{j} \Delta x + \dfrac{n-1}{2} \times d, y_j - c\right) \end{cases} \quad (2)$$

where $\sum\limits_{j}^{j+c}$ denotes the summation along the positive semi-axis of the *y*-axis, and $\sum\limits_{j-c}^{j}$ represents the summation along the negative semi-axis of the *y*-axis.



(**a**)

(**b**)

**Figure 4.** Schematic diagram of the convolution kernel computed along the x- and y-axis directions. (**a**) *x*-axis; (**b**) *y*-axis.

According to Equations (1) and (2), an example of the selectable range of receptive fields for the dynamic dilated serpentine convolution kernel during the feature extraction process is illustrated in Figure 5.

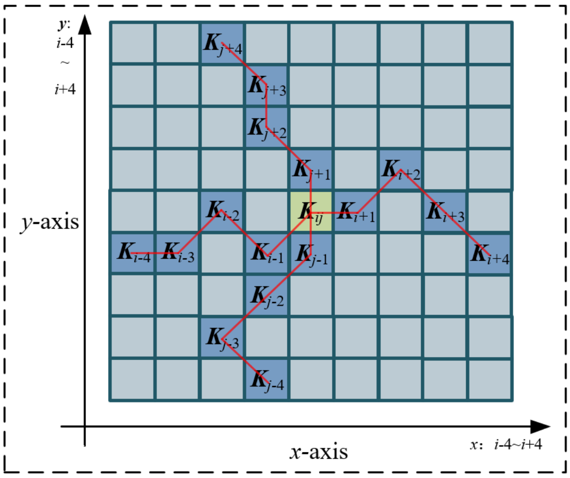


**Figure 5.** DDSConv feeling the range of wild options.

The Dynamic Dilated Snake Convolutional Layer (DDSConv Layer) based on a 3×3 standard convolution architecture is illustrated in Figure 6. This structure employs two dilation rate parameters (1 and 2) to expand the receptive field. To enhance adaptive feature representation, the design systematically integrates a Squeeze-and-Excitation (SE) module and achieves weighted fusion of multi-level features through cross-layer connections. The DDSConv Layer dynamically optimizes the morphological configuration and dilation parameters of the convolution kernel according to input feature characteristics. This dynamic tunability enables the network to capture multi-scale spatial patterns with higher precision, significantly improving the granularity of feature analysis, and demonstrates superior performance in aerial imagery object detection tasks requiring robust handling of complex background interference.

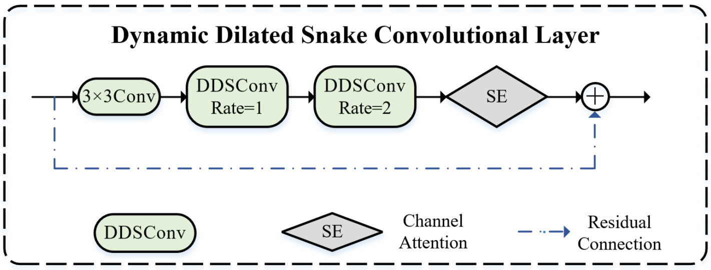


**Figure 6.** DDSConv Layer structure.

### 3.2.3. Multi-Scale Feature Aggregation Module (MFAM)

To simultaneously capture deep and shallow spatial features across different hierarchy levels and enable convolutional kernels to adapt to varying contextual environments, this paper proposes a MFAM. By introducing a spatial attention mechanism, the design enhances the kernel's adaptive capability to detect region-specific and target-relevant features at different spatial positions, thereby comprehensively considering global and local features with diverse sizes and shapes across diverse samples. The module incorporates two independent spatial attention units dedicated to processing global and local features respectively, as shown in Figure 7.

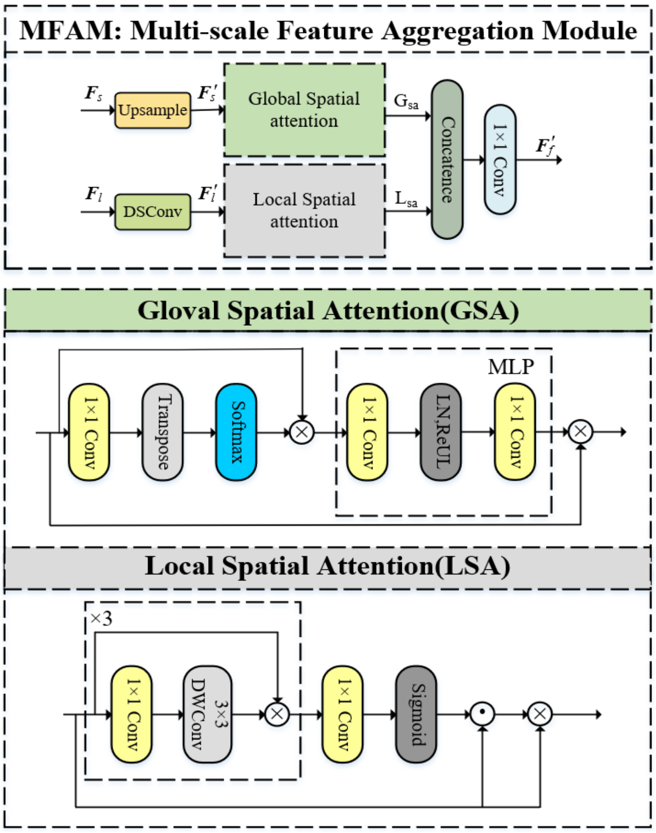


**Figure 7.** MFAM structure.

Since shallow feature maps preserve richer local characteristics of multi-scale targets, the global spatial attention unit employs 1×1 small convolutional kernels to meticulously capture global features from shallow layers. In contrast, the local spatial attention unit utilizes 3×3 large convolutional kernels to extract global patterns from deep feature maps. This dual-branch architecture effectively preserves localized details and global contextual modeling, while separated channel dimensions optimize the balance between detection accuracy and computational resources. Finally, the outputs of both attention units are concatenated and processed through a 1×1 convolutional layer, with the mathematical formulation of this operation defined as:

$$F_f^1, F_f^2 = Split\left(F_f\right) \qquad (3)$$

$$F_f' = pwConv\left(Concat\left(G_{sa}\left(F_f^1\right), L_{sa}\left(F_f^2\right)\right)\right) \qquad (4)$$

where $Split(\bullet)$ denotes the feature map split operation. $Concat$ represents the output concatenation operation. $pwConv$ corresponds to the pointwise convolution using a 1×1 kernel. $G_{sa}$ denotes the global spatial attention and $L_{sa}$ indicates the local spatial attention.

The GSA module focuses on long-range dependencies between pixels, serving as a complementary mechanism to local spatial attention. By capturing these long-range interactions, it significantly enhances the representational capacity of features. Let X be the input feature map. The process for generating global spatial attention is defined as follows:

$$Att_G\left(F_f^1\right) = Softmax\left(Transpose\left(pwConv\left(F_f^1\right)\right)\right) \qquad (5)$$

$$G_{sa}\left(F_f^1\right) = MLP\left(Att_G\left(F_f^1\right) \otimes F_f^1\right) + F_f^1 \qquad (6)$$

where $Att_G(\bullet)$ denotes the global attention operator. $Softmax(\bullet)$ represents the Softmax activation function. $Transpose$ corresponds to the transpose operation. $MLP$ is composed of two pointwise convolution layers, a ReLU nonlinear activation function, and a fully connected layer. $\otimes$ indicates the matrix multiplication operation.

The LSA module focuses more on local features within the spatial dimensions of a given feature map. Using the sub-feature map $F_i^2$ as input, the calculation formula is defined as:

$$Att_L\left(F_f^2\right) = Sigmoid\left(pwConv\left(F_c\left(F_f^2\right) + F_f^2\right)\right) (7)$$

$$L_{sa}\left(F_f^1\right) = Att_L\left(F_f^2\right) \odot F_f^2 + F_f^2 \qquad (8)$$

where $Att_L(\bullet)$ denotes the local attention operator. $Sigmoid(\bullet)$ represents the Sigmoid activation function. $F_c(\bullet)$ consists of three stacked 1×1 convolutional layers and a 3×3 depthwise separable convolution layer. $\odot$ indicates element-wise matrix multiplication. This structural design efficiently focuses on local spatial information with fewer parameters.

### 3.2.4. MFAM-Neck

In drone-captured images, target scale variations are often highly significant, with a prevalence of small-sized targets. To address these challenges in target detection, this paper proposes a novel MFAM-Neck framework, designed to adapt to hierarchical contextual environments. This framework effectively integrates shallow-layer features and deep-layer features from multi-scale feature maps. By leveraging the MFAM module to reconstruct fused feature information, MFAM-Neck provides high-quality feature support for target detection. The detailed architecture of the MFAM-Neck network is illustrated in Figure 8.

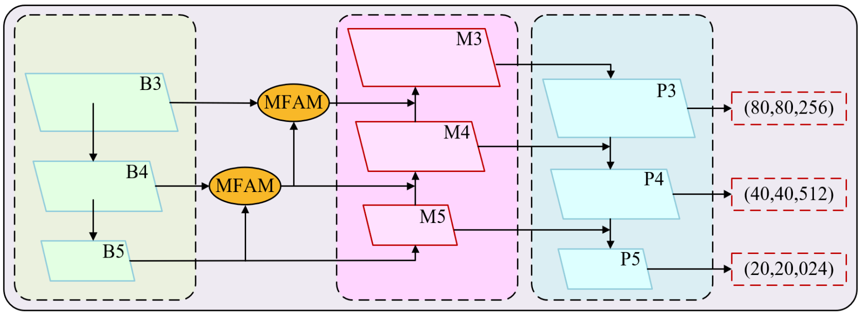


**Figure 8.** MFAM-Neck structure.

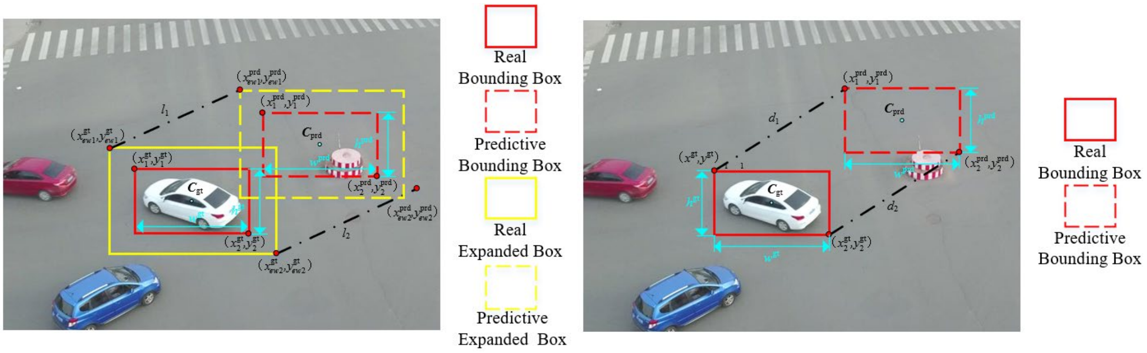
The MFAM-Neck feature fusion network adopts a multi-scale fusion strategy, enhancing feature representation by integrating B3, B4, and B5 feature maps from different hierarchical layers of the backbone network. Given their distinct receptive fields and semantic abstraction levels, the network incorporates a MFAM in the Neck section. This module achieves effective feature fusion across dual-scale spatial features through an adaptive weight allocation mechanism. Specifically, the network employs bidirectional feature propagation paths, including top-down feature propagation and bottom-up feature aggregation. After multi-level feature fusion, the final output feature maps P3, P4, and P5 have dimensions of 80×80×256, 40×40×512, and 20×20×1024, respectively, reflecting multi-scale feature representations.

### 3.2.5. Loss Functions

The complex structures, background textures, and noise in drone aerial images often interfere with target localization and recognition, thereby degrading detection accuracy. By rationally designing the loss function, model convergence can be accelerated and regression performance improved. YOLOv8 introduces Distribution Focal Loss (DFL) and CIoU to evaluate bounding box regression losses. Although CIoU accounts for the center distance and aspect ratio between boxes, its aspect ratio is defined as a relative value rather than an absolute value, leading to suboptimal balancing of difficulty levels across different samples. Additionally, the use of inverse trigonometric functions in its calculation may increase computational overhead.

To address these issues and further accelerate model convergence, this paper introduces the concept of Inner-IoU, which employs auxiliary bounding boxes to enhance convergence, and proposes an Expanded-Window Bounding Box Regression Loss Function (EW-BBRLF). The core idea of EW-BBRLF is to improve bounding box regression through anchor box expansion and minimum point distance optimization. Building on the strengths of CIoU and Ln-norm loss, EW-BBRLF introduces a scaling factor to control the size of auxiliary bounding boxes, thereby enhancing localization precision and detection accuracy.

EW-BBRLF employs larger auxiliary bounding boxes when calculating IoU loss, effectively promoting regression for low-IoU samples and accelerating model convergence. Conversely, smaller auxiliary boxes aid in precise localization of high-IoU samples. However, small aerial targets often exhibit low IoU values due to their tiny size and insufficient feature information. Thus, leveraging large-scale expanded windows significantly improves bounding box regression accuracy for small targets. A comparative visualization of EW-BBRLF and CIoU is shown in Figure 9.



(**a**) (**b**)

**Figure 9.** Comparison of EW-BBRLF and CIoU. (**a**) EW-BBRLF; (**b**) CIoU.

Given the coordinates of the predictive bounding box

$$\boldsymbol{B}_{\text{prd}} = \left(x_1^{\text{prd}}, y_1^{\text{prd}}, x_2^{\text{prd}}, y_2^{\text{prd}}\right) \qquad (9)$$

coordinates of the real bounding box

$$\boldsymbol{B}_{\text{gt}} = \left(x_1^{\text{gt}}, y_1^{\text{gt}}, x_2^{\text{gt}}, y_2^{\text{gt}}\right) \qquad (10)$$

coordinates of the centre of the predictive bounding box

$$\boldsymbol{C}_{\mathrm{prd}} = \left(x_c^{\mathrm{prd}}, y_c^{\mathrm{prd}}\right) \qquad (11)$$

coordinates of the centre of the real bounding box

$$\boldsymbol{C}_{\mathrm{gt}} = \left(x_c^{\mathrm{gt}}, y_c^{\mathrm{gt}}\right) \, (12)$$

predictive box width, height

$$\boldsymbol{L}_{\mathrm{prd}} = \left(w^{\mathrm{prd}}, h^{\mathrm{prd}}\right) \qquad (13)$$

real box width, height

$$\boldsymbol{L}_{\mathrm{gt}} = \left(w^{\mathrm{gt}}, h^{\mathrm{gt}}\right) \, (14)$$

then the predicted expanded box is

$$\boldsymbol{A}_{\mathrm{prd}} = \left(x_{ew1}^{\mathrm{prd}}, y_{ew1}^{\mathrm{prd}}, x_{ew2}^{\mathrm{prd}}, y_{ew2}^{\mathrm{prd}}\right) \qquad (15)$$

where, $(x_{ew1}^{\mathrm{prd}}, y_{ew1}^{\mathrm{prd}})$ and $(x_{ew2}^{\mathrm{prd}}, y_{ew2}^{\mathrm{prd}})$ represent the top-left corner coordinate and bottom-right corner coordinate of the predicted expanded window, respectively. The calculation formula is defined as:

$$x_{ew1}^{\mathrm{prd}} = x_1^{\mathrm{prd}} - \frac{w^{\mathrm{prd}} \cdot r}{2} \qquad (16)$$

$$y_{ew1}^{\mathrm{prd}} = y_1^{\mathrm{prd}} - \frac{h^{\mathrm{prd}} \cdot r}{2} \qquad (17)$$

$$x_{ew2}^{\mathrm{prd}} = x_2^{\mathrm{prd}} + \frac{w^{\mathrm{prd}} \cdot r}{2} \qquad (18)$$

$$y_{ew2}^{\mathrm{prd}} = y_2^{\mathrm{prd}} + \frac{h^{\mathrm{prd}} \cdot r}{2} \qquad (19)$$

where $r$ is the scaling factor to control the size of the expanded window, which is set to 1.2. Similarly, the true expanded window can be derived as:

$$\boldsymbol{A}_{\mathrm{gt}} = \left(x_{ew1}^{\mathrm{gt}}, y_{ew1}^{\mathrm{gt}}, x_{ew2}^{\mathrm{gt}}, y_{ew2}^{\mathrm{gt}}\right) \qquad (20)$$

The IoU of the bounding boxes is

$$IoU = \frac{inter}{w^{gt} h^{gt} + w^{prd} h^{prd} - inter} \qquad (21)$$

where $inter$ denotes the overlapping area between the predicted bounding box and ground truth bounding box, calculated as

$$inter = \left(\min\left(x_2^{prd}, x_2^{gt}\right) - \max\left(x_1^{prd}, x_1^{gt}\right)\right) \cdot \left(\min\left(y_1^{prd}, y_1^{gt}\right) - \max\left(y_2^{prd}, y_2^{gt}\right)\right) \qquad (22)$$

Similarly, the IoU of the expanded window is

$$IoU^{EW} = \frac{inter}{\left(w^{\mathrm{gt}} h^{\mathrm{gt}} + w^{\mathrm{prd}} h^{\mathrm{prd}}\right) \mathrm{ratio}^2\text{-}inter} \qquad (23)$$

The minimum-point-distance-based expanded window IoU is

$$IoU^{C} = IoU - \frac{l_1^2}{w^2 + h^2} - \frac{l_2^2}{w^2 + h^2} \qquad (24)$$

$$l_1^2 = \left(x_{ew1}^{\mathrm{prd}} - x_{ew1}^{\mathrm{gt}}\right)^2 + \left(y_{ew2}^{\mathrm{prd}} - y_{ew2}^{\mathrm{gt}}\right)^2 \qquad (25)$$

$$l_2^2 = \left(y_{ew1}^{\mathrm{prd}} - y_{ew1}^{\mathrm{gt}}\right)^2 + \left(x_{ew2}^{\mathrm{prd}} - x_{ew2}^{\mathrm{gt}}\right)^2 \qquad (26)$$

where $w$ and $h$ represent the width and height of the image, respectively.

The loss function incorporating the expanded window is

$$L_C = 1 - IoU^C \quad (27)$$

Finally, the EW-BBRLF calculation formula is defined as

$$L_{EW-BRRLF} = L_C + IoU - IoU^{EW} \, (28)$$

EW-BBRLF comprehensively integrates factors including aspect ratio differences, center point distance, and overlapping regions, while streamlining computational processes. This effectively addresses limitations inherent in the CIoU method. Furthermore, by controlling the scaling factor to generate larger expanded windows, EW-BBRLF mitigates challenges arising from small target sizes and insufficient feature information, significantly enhancing detection performance for small targets.

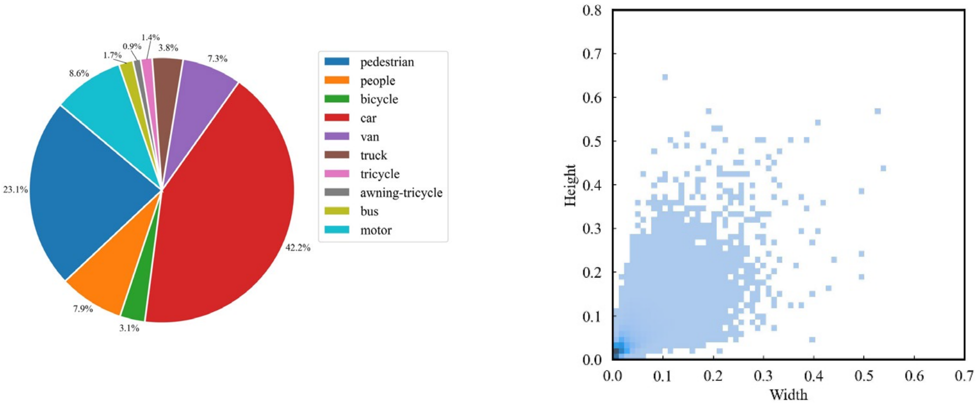
doi:10.20944/preprints202506.0837.v1

## 4. Experimental Results and Analysis

### *4.1. Datasets*

To validate the effectiveness of the proposed DMF-YOLO model, comparative experiments were conducted using the publicly available VisDrone dataset [47], with additional validation of the model's generalization capability performed on the infrared dataset InfraredData. The VisDrone dataset, collected by Tianjin University using UAVs under diverse low-altitude conditions, contains images in two resolutions: 1360×765 and 960×540 pixels. It encompasses various weather conditions, illumination scenarios, and sparsity levels in daily life scenes, covering 10 object categories: Pedestrian (Ped), Person (Peo), Bicycle (Bic), Car, Van, Truck (Tru), Tricycle (Tri), Awning-tricycle (Awn), Bus, and Motorcycle (Mot). Specifically, the dataset includes 6,471 training images, 3,190 test images, and 548 validation images. Figure 10 illustrates the distribution of label categories and object sizes. The detection challenges of this dataset mainly manifest in the following aspects: images contain a large number of objects, particularly small and extremely small targets, coupled with severe occlusion between objects and imbalanced data distribution. These factors collectively increase detection complexity, yet the dataset's richness and diversity provide a solid foundation for effectively evaluating model performance.

In contrast, the HIT-UAV [48] infrared object detection dataset released by Harbin Institute of Technology consists of infrared aerial data captured by drones, primarily designed for detecting humans and vehicles. This dataset contains 2,898 infrared thermal imaging images collected by UAVs. The images cover diverse scenarios including campuses, parking areas, roads, and sports fields, and encompass multiple target categories (pedestrians, bicycles, cars, and other vehicles). Additionally, the data incorporates varied acquisition parameters such as flight altitudes of 60-130 meters, camera tilt angles ranging from 30° to 90°, and diverse lighting conditions across different time periods. Figure 11 displays the category label types and size distribution within the infrared dataset.



(**a**) (**b**)

**Figure 10.** Object type and size distribution of the VisDrone dataset. (**a**) Types of objects; (**b**) Object size distribution.

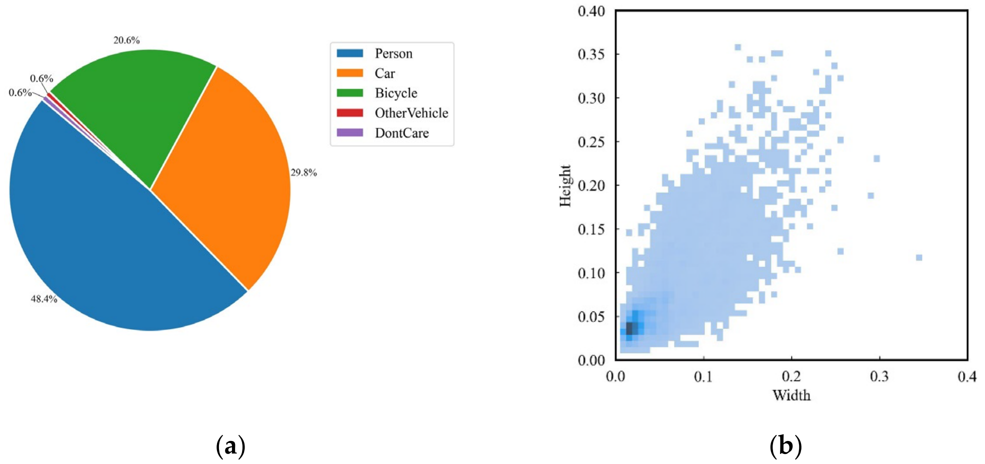


**Figure 10.** Object type and size distribution of the HIT-UAV dataset. (**a**) Types of objects; (**b**) Object size distribution.

As shown in Figures 10(b) and 11(b), the horizontal and vertical axes represent the width and height of bounding boxes, respectively, illustrating the size distribution of bounding boxes in both datasets. Analysis reveals that both datasets encompass multi-scale objects, with a notably high proportion of small-sized targets. This distribution pattern fully reflects the critical challenges inherent in object detection tasks for UAV aerial imagery, which closely aligns with the core research focus addressed in this paper.

### *4.2. Experimental Environment and Parameters*

This study was conducted on a Windows 10 operating system using PyTorch 1.11.8 as the deep learning framework, with Python 3.8 as the compiler and CUDA 11.8 for GPU acceleration. All experiments were trained, validated, and tested on an NVIDIA RTX 3090 GPU. The hyperparameters used during training are listed in Table 1.

**Table 1.** Training Hyperparameters.

| Name | Value | Name | Value |
|---|---|---|---|
| Optimizer | SGD | Training Epochs | 150 |
| Image Size | 640×640;512×512 | Workers | 16 |
| Initial Learning Rate | 0.01 | Learning Rate Decay | 0.0001 |
| Weight Decay | 0.0005 | Batch Size | 8 |
| Momentum Factor | 0.937 | Warmup Epochs | 3 |

### *4.3. Evaluation Metrics*

To evaluate the small object detection performance of DMF-YOLO, we employ Precision (P), Recall (R), Mean Average Precision (mAP), and Parameter Count (Par) as evaluation metrics.

(1) Precision (P) represents the ratio of correctly detected objects to the total number of detections, reflecting the model's accuracy. The formula is:

$$\mathrm{P} = \frac{TP}{TP + FP} \quad (29)$$

where $TP$ is the sample that was predicted to be positive and actually positive, and $FP$ is the sample that was predicted to be positive and actually negative.

(2) Recall (R) represents the ratio of correctly detected objects to the total number of ground-truth objects, reflecting the model's detection coverage. The formula is:

$$\mathrm{R} = \frac{TP}{TP+FN} \quad (30)$$

(3) Average Precision (AP) is defined as the area under the Precision-Recall (P-R) curve. Specifically, an Intersection over Union (IoU) threshold is first set. Using the recall rate corresponding to this threshold as the x-axis and precision as the y-axis, the P-R curve is plotted. AP is calculated by averaging precision values along the P-R curve. The formula for AP is:

$$\mathrm{AP} = \int_0^1 \mathrm{P} \cdot \mathrm{RdR} \qquad (31)$$

Mean Average Precision (mAP) is obtained by computing the weighted average of AP values across all object categories, which evaluates the model's detection performance on all classes. The formula is:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i \qquad (32)$$

where $\mathrm{AP}_i$ denotes the AP value for the $i$-th category, and $N$ represents the total number of object categories in the training dataset.

The mAP metric comprehensively reflects the model's overall detection performance across all categories. Here, mAP50 refers to the mean AP across all categories when the IoU threshold is set to 0.5, while mAP50:95 averages detection accuracy over 10 IoU thresholds ranging from 0.5 to 0.95 (with a step size of 0.05). It is noteworthy that higher IoU thresholds impose stricter requirements on the model's detection capability. The IoU is defined as the ratio of the intersection area to the union area between the predicted bounding box and the ground-truth annotation.

(4) Parameter Count (Par): The total number of trainable parameters in the model during training.

## *4.4. Ablation Study Analysis*

To validate the effectiveness of the proposed improvements and their contributions to model performance, this study selects YOLOv10s as the baseline model on the VisDrone dataset and employs metrics including Precision (P), Recall (R), mAP50, mAP50:95, and Par for comprehensive evaluation. A series of ablation experiments with different combinations of improvement modules were conducted to analyze the detection effects of each proposed method. During this process, P and R were calculated under an IoU threshold of 0.5 and a confidence threshold of 0.3.

### 4.4.1. Effectiveness Analysis of Single and Multi-Module Improvements

The ablation results for single-module and multi-module improvements are shown in Tables 2 and 3, respectively. Here, DDSConv denotes Dynamic Dilated Snake Convolution, MFAM represents the Multi-scale Feature Aggregation Module, EW-BBRLF indicates the Expanded Window-based Bounding Box Regression Loss Function, and √ indicates the adoption of this improvement strategy.

**Table 2.** Single-Module Ablation Experiments.

| **DDSConv** | **MFAM** | **EW-BBRLF** | **P**(%) | **R**(%) | **mAP50**(%) | **mAP50:95**(%) | **Par**(M) |
|---|---|---|---|---|---|---|---|
| - | - | - | 49.3 | 38.6 | 39.5 | 22.8 | 13.3 |
| √ | | | 52.3 | 39.4 | 42.0 | 23.1 | 15.1 |
| | √ | | 54.0 | 42.2 | 44.3 | 25.4 | 18.6 |
| | | √ | 49.8 | 39.7 | 40.6 | 23.4 | 11.3 |

**Table 3.** Multi-Module Ablation Experiments.

| DDSConv | MFAM | EW-BBRLF | P(%) | R(%) | mAP50(%) | mAP50:95(%) | Par(M) |
|---|---|---|---|---|---|---|---|
| - | - | - | 49.3 | 38.6 | 39.5 | 22.8 | 13.3 |
| √ | | | 52.3 | 39.4 | 42.0 | 23.1 | 15.1 |
| √ | √ | | 58.2 | 46.1 | 48.3 | 30.4 | 21.6 |
| √ | | √ | 54.6 | 43.2 | 45.8 | 27.4 | 15.1 |
| | √ | √ | 57.9 | 44.2 | 46.5 | 28.3 | 18.6 |
| √ | √ | √ | 60.4 | 48.6 | 51.9 | 31.7 | 21.6 |

From the single-module ablation experiments in Table 2, DDSConv expands the receptive field and learns richer feature information, thereby better adapting to the detection requirements for tiny objects and complex shape variations in aerial images. By incorporating domain knowledge about micro-structural morphology during feature extraction, it stably enhances the perception of slender columnar targets. This results in increases of 6.1% in P, 2.1% in R, 6.3% in mAP50, and 1.3% in mAP50:95, with only a 16% rise in parameter count, demonstrating the module's effectiveness. MFAM introduces a spatial attention mechanism to strengthen the convolution kernels' adaptive capability for detecting location-specific and small-object-related features, achieving significant boosts of 12.1% in both mAP50 and mAP50:95, proving its efficacy in extracting small-target features. EW-BBRLF enhances bounding box localization accuracy through anchor expansion and minimum-point-distance-based regression, improving mAP50 by 2.8%.

Table 3 presents multi-module ablation experiments, validating the effectiveness of combined improvements. The integration of DDSConv and MFAM reduces small-object information loss and suppresses redundant noise during feature extraction, further enhancing DDSConv's ability to fuse more effective information, thereby increasing mAP50 by 15%. The combination of DDSConv and EW-BBRLF achieves optimal lightweight performance with only a 16% parameter increase, but due to insufficient multi-scale feature fusion, mAP50 improves by only 11.9%. The Improvement MFAM and EW-BBRLF combination fails to effectively mitigate small-object information loss during feature extraction. In contrast, the joint application of Improvements 1, 2, and 3 successfully addresses small-object detection and multi-scale processing challenges, achieving a remarkable 31.4% improvement in mAP50. The ablation study results confirm that the proposed methods significantly enhance small-object detection performance, fully validating the effectiveness of the designed algorithms.

### 4.4.2. Effectiveness of the MFAM-Neck Module

This study proposes a Neck network architecture based on the MFAM module. To evaluate the performance advantages of MFAM, we designed three MFAM-Neck variants with different feature layer fusion strategies, named MFAM-Neck-A, MFAM-Neck-B, and MFAM-Neck-C, respectively. Their detailed architectures are illustrated in Figure 12. Experiments compare the performance of optimized backbone networks, with specific results listed in Table 4.

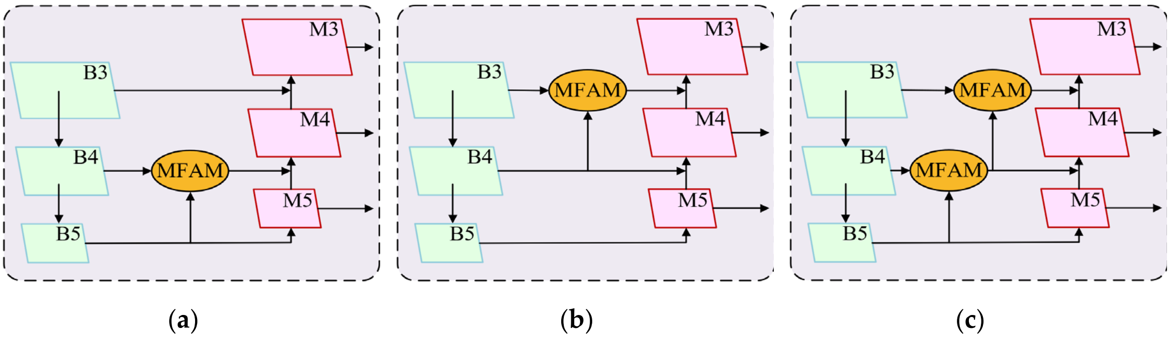


**Figure 12.** Detailed Structures of the Three MFAM-Neck Variants. (**a**) MFAM-Neck-A; (**b**) MFAM-Neck-B; (**c**) MFAM-Neck-C.

**Table 4.** Experimental Comparison Results of the Three MFAM-Neck Variants.

| **Method** | **mAP50**(%) | **mAP50:95**(%) | **Par**(M) |
|---|---|---|---|
| YOLOv10s | 39.5 | 22.8 | 11.3 |
| YOLOv10l | 44.5 | 25.7 | 30.5 |
| MFAM-Neck-A | 42.3 | 24.5 | 15.8 |
| MFAM-Neck-B | 42.9 | 24.9 | 15.6 |
| MFAM-Neck-C | 44.3 | 25.4 | 18.6 |

As shown in Table 4, compared to the baseline algorithm YOLOv10s, all three MFAM-Neck architectures significantly improve object detection accuracy. Specifically, mAP50 increases by 7.1%, 8.6%, and 12.2%, while mAP50:95 improves by 7.5%, 9.2%, and 11.4%, respectively. All three improved architectures integrate the MFAM module to achieve multi-level feature fusion, thereby enhancing the model's feature representation capability. Notably, the MFAM-Neck-C architecture achieves detection accuracy comparable to YOLOv10m but with a substantially reduced parameter count. Experimental results indicate that among the three variants, MFAM-Neck-C exhibits the best detection performance. Although its mAP50 and mAP50:95 metrics are 0.4% and 1.2% lower than those of YOLOv10m, respectively, this architecture reduces the parameter count by 39.1%. This result demonstrates that the MFAM-Neck structure achieves near-large-scale model accuracy at the cost of minimal complexity increases. The performance improvement primarily benefits from the multi-layer feature fusion strategy of MFAM-Neck, which enriches feature representations for multi-scale targets by integrating semantic information from different hierarchical levels.

### *4.5. Experimental Results and Analysis on VisDrone Test Set*

To evaluate the performance of DMF-YOLO, this study conducts comparative analyses with several representative aerial image object detection methods: RetinaNet [49], CenterNet [50], QueryDet [31], YOLOv5s, YOLOv8s, MCA-YOLOv5 [51], YOLOv10s, DAMO-YOLOv10 [52], and CA-YOLO [53]. Table 5 presents the comparative experimental results of these algorithms on the VisDrone-2019 dataset under identical experimental conditions, where bold numbers indicate optimal values. Additionally, to comprehensively assess DMF-YOLO's detection performance across diverse scenarios, five typical scene categories were selected for comparative experiments: dense scenarios, complex background scenarios, motion blur scenarios, and long-distance target scenarios. Cross-scenario object detection results are illustrated in Figure 13, while heatmap comparisons for dense and long-distance target scenarios are shown in Figures 14 and 15.

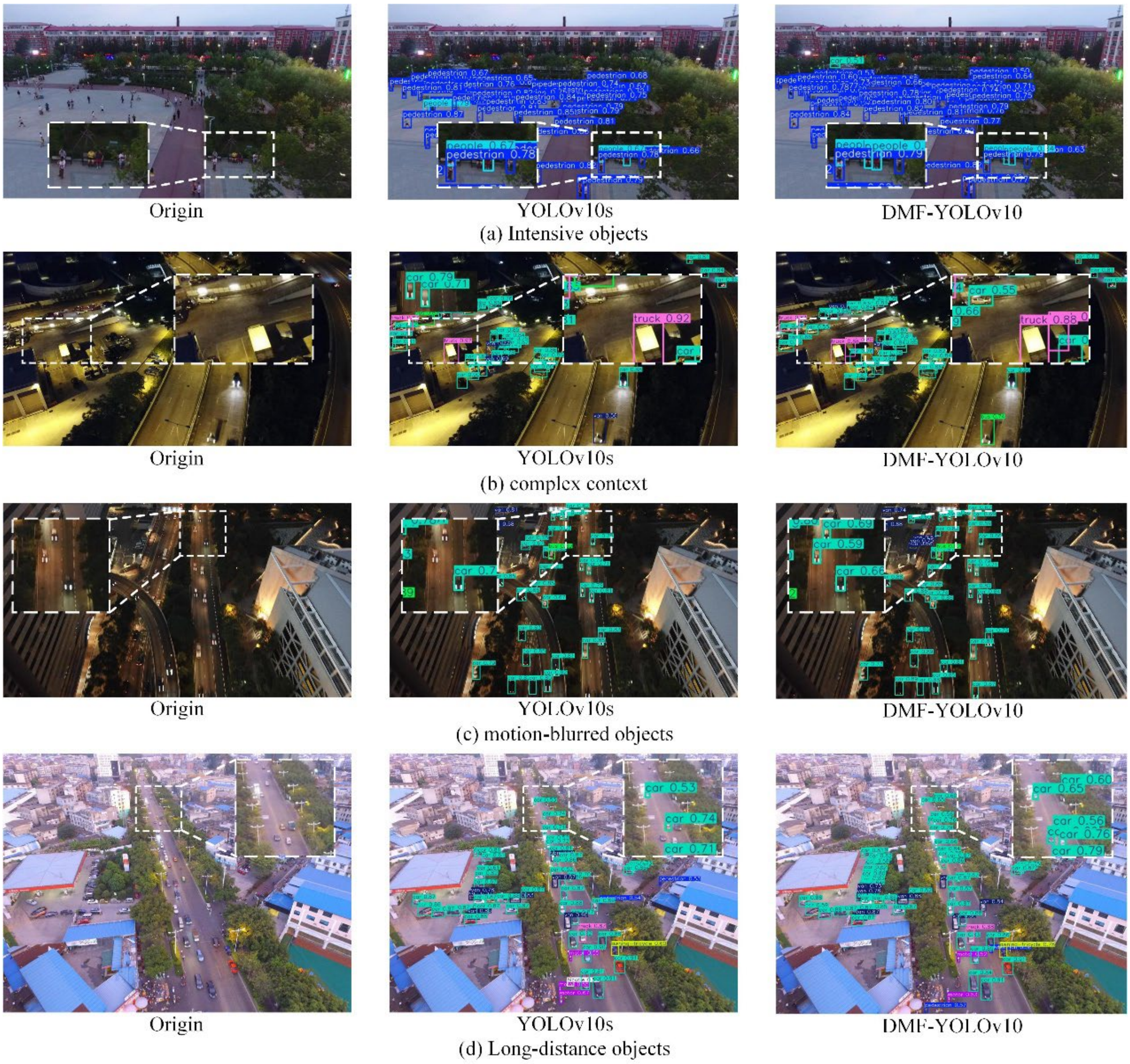
**Table 5.** Comparative Experimental Results of Different Methods.

| **Method** | **AP**(%) | | | | | | | | | | **m1**(%) | **P**(%) | **R**(%) | **m2**(%) | **Par**(M) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ped | Peo | Bic | Car | Van | Tru | Tri | Awn | Bus | Mot | all | | | | |
| RetinaNet | 13.0 | 7.9 | 1.4 | 45.5 | 19.9 | 11.5 | 6.3 | 4.2 | 17.8 | 11.8 | 13.9 | 37.5 | 28.4 | 12.0 | 15.8 |
| CenterNet | 22.6 | 20.6 | 14.6 | 59.7 | 24.0 | 21.3 | 20.1 | 17.4 | 37.9 | 23.7 | 26.2 | 39.8 | 30.3 | 14.3 | 19.1 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| QueryDet | 56.8 | 37.4 | 17.6 | 80.3 | 41.9 | 41.8 | 24.2 | 10.1 | 62.1 | 44.8 | 41.7 | 48.3 | 37.1 | 21.2 | **35.6** |
| YOLOv5s | 37.9 | 31.5 | 12.8 | 70.4 | 34.2 | 31.7 | 18.7 | 12.6 | 41.1 | 37.2 | 32.8 | 42.4 | 33.6 | 17.6 | 7.2 |
| MCA-YOLOv5 | 42.0 | 27.8 | 18.1 | 81.6 | 47.1 | **52.0** | 26.8 | 25.3 | 63.2 | 43.1 | 42.7 | - | - | 28.3 | 31.8 |
| YOLOv8s | 42.7 | 32.6 | 14.2 | 80.6 | 44.4 | 46.8 | 28.2 | 23.1 | 55.7 | 44.5 | 41.2 | 49.6 | 41.7 | 23.1 | 18.6 |
| YOLOv10s | 43.2 | 32.6 | 12.9 | 79.5 | 46.1 | 36.3 | 28.6 | 15.6 | 54.8 | 45.4 | 39.5 | 49.3 | 38.6 | 22.8 | 13.3 |
| DAMO-YOLOv10 | 50.8 | 43.8 | **26.0** | 81.1 | 53.7 | 49.9 | **41.8** | **27.9** | 67.1 | 53.3 | 47.5 | 55.9 | 43.2 | 25.4 | 16.4 |
| CA-YOLO | 55.5 | 45.8 | 23.5 | 85.5 | 52.7 | 42.1 | 38.2 | 22.3 | 64.6 | 57.5 | 48.8 | **62.1** | 45.1 | 27.6 | 31.1 |
| ours | **56.6** | 47.4 | 24.6 | **87.5** | **53.1** | 41.7 | 39.8 | 25.4 | **65.9** | **59.7** | **50.1** | 60.4 | **48.6** | **29.7** | 17.6 |

Note: m1 is the mAP50 indicator and m2 is the mAP50:95 indicator.

As shown in Table 5, DMF-YOLO demonstrates superior detection accuracy compared to other algorithms on the VisDrone-2019 dataset, which is dominated by small targets. While traditional object detection methods like RetinaNet and CornerNet hold theoretical significance, their significantly higher parameter counts make their detection efficiency inadequate for real-time UAV image processing. In contrast, the YOLO series algorithms maintain high detection accuracy while exhibiting superior real-time performance, making them more suitable for UAV-based detection tasks. Compared to earlier YOLO variants, YOLOv10 shows notable improvements in both detection accuracy and inference speed. Notably, although YOLOv10s exhibits an 8.5% reduction in mAP compared to YOLOv8s, its parameter count is dramatically reduced by 28.4% (from 18.6M for YOLOv8s to 13.3M for YOLOv10s), highlighting its exceptional balance between model efficiency and detection performance. Building upon YOLOv10s as the baseline, this work introduces Dynamic Dilated Snake Convolution into the backbone network to enhance local feature extraction and representation for small targets. Additionally, the MFAM-Neck further integrates multi-scale features, significantly improving the model's capability to detect small targets in aerial images. Experimental results show that at an input resolution of 640×640 pixels, the proposed algorithm achieves outstanding detection performance for common vehicle categories: AP50 scores for cars, vans, and buses reach 87.5%, 53.1%, and 65.9%, respectively. Other comparison models fail to effectively capture structural features of small targets or leverage high-resolution low-level feature maps for fusion, resulting in substantial loss of fine-grained details. In contrast, DMF-YOLO successfully addresses these limitations, achieving mAP50 and mAP50:95 scores of 50.1% and 29.7%, respectively, with a parameter count of only 17.6M. This performance advantage confirms that the algorithm achieves an ideal balance between accuracy and efficiency, making it particularly suitable for UAV aerial image object detection applications.



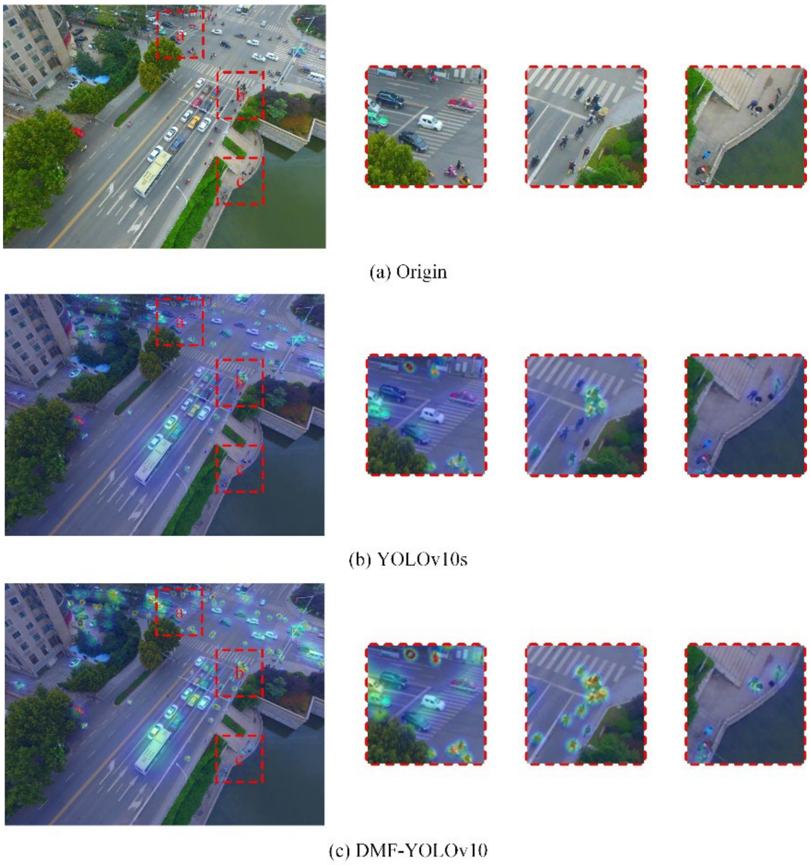
(a) Intensive objects

(b) complex context

(c) motion-blurred objects

(d) Long-distance objects

**Figure 13.** Comparison of object detection in different scenes.

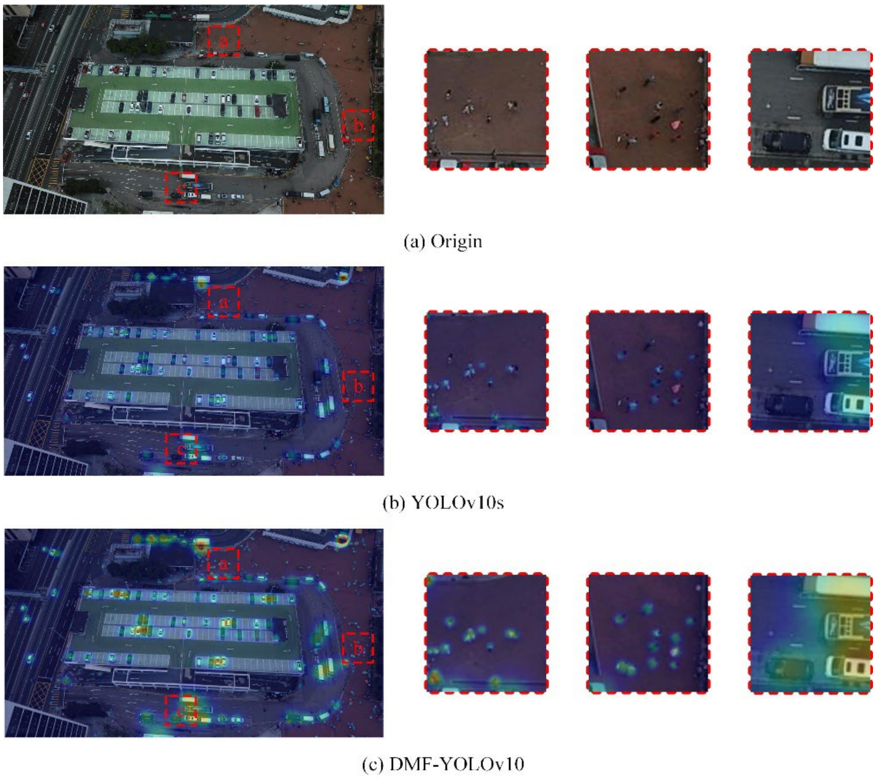


(a) Origin

(b) YOLOv10s

(c) DMF-YOLOv10

**Figure 14.** Comparison of object detection in different scenes.



(a) Origin

(b) YOLOv10s

(c) DMF-YOLOv10

**Figure 15.** Comparison of heat maps in scenes with distant objects.

The comparative experiments in Figure 13 reveal limitations in the baseline model YOLOv10s across diverse complex scenarios: miss detection of individuals seated on chairs in dense scenes, failure to detect vehicles parked near walls in complex background scenarios, omission of high-speed vehicles in motion blur scenarios, and inability to identify certain long-distance targets in remote scenes. In stark contrast, the improved DMF-YOLOv10 model demonstrates significant optimization in these challenging scenarios, particularly showing enhanced performance in detecting small targets such as pedestrians and distant vehicles.

The heatmap comparisons in Figures 14 and 15 further elucidate the improvement mechanism: the original model exhibits insufficient attention to distant vehicles and small-sized pedestrians in dense scenes, and lacks effective response to most pedestrians and black cars blending with the background in high-altitude small-target scenarios. DMF-YOLOv10 achieves breakthroughs through two innovative mechanisms:

- Dynamic Dilated Snake Convolution kernels are introduced during feature extraction, substantially enhancing the network's capability to capture critical features.
- A novel multi-scale feature fusion strategy is implemented by integrating high-resolution details from large-scale feature maps with deep semantic information from small-scale feature maps. This dual optimization simultaneously improves sensitivity to small targets and robustness for detecting background-similar objects.

This two-dimensional enhancement enables DMF-YOLOv10 to exhibit superior detection accuracy compared to the baseline model, particularly demonstrating significant advantages in handling UAV aerial image-specific scenarios characterized by complex small targets.

### *4.6. Experimental Results and Analysis of the HIT-UAV Test Set*

To evaluate the generalization capability of the proposed algorithm, this study conducted cross-dataset validation experiments on the HIT-UAV dataset and performed systematic comparisons with other object detection models. Table 6 presents the test results of different models under identical experimental conditions, while Figure 16 illustrates the mAP50 comparisons across different categories. Specific detection results under various scenarios are visualized in Figure 17.

**Table 6.** Test results of different models.

| **Method** | **mAP50**(%) | **mAP50:95**(%) | **Par**(M) | **Test Speed** (ms) |
|---|---|---|---|---|
| YOLOv5s | 77.2 | 47.9 | 15.6 | 4.8 |
| YOLOv8s | 80.1 | 49.3 | 18.2 | 5.6 |
| YOLOv10s | 79.3 | 49.5 | **12.5** | **3.8** |
| YOLOv10l | 80.6 | 50.9 | 43.2 | 13.6 |
| CA-YOLOv10 | 81.1 | 51.4 | 23.5 | 8.9 |
| Ours | **81.4** | **52.8** | 14.2 | 4.5 |

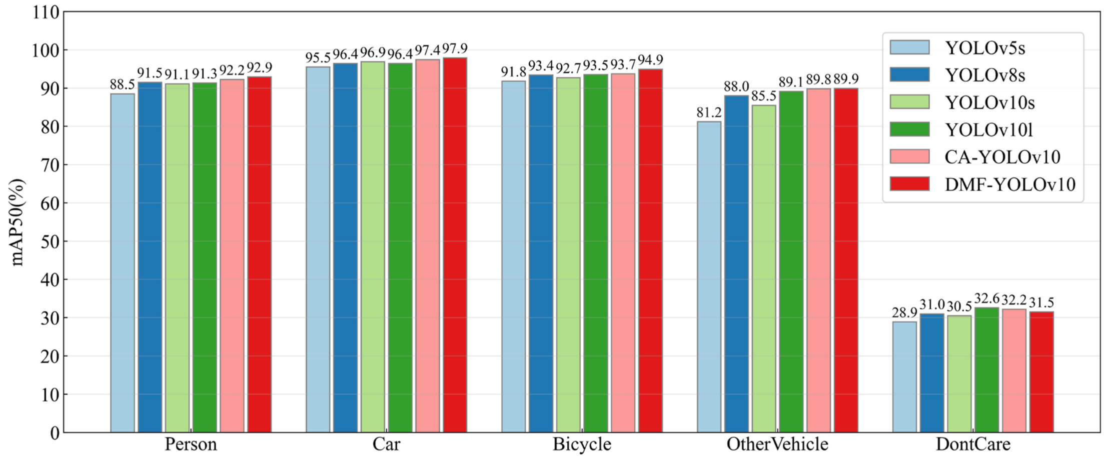


**Figure 16.** mAP50 Comparisons of Different Models Across Categories.

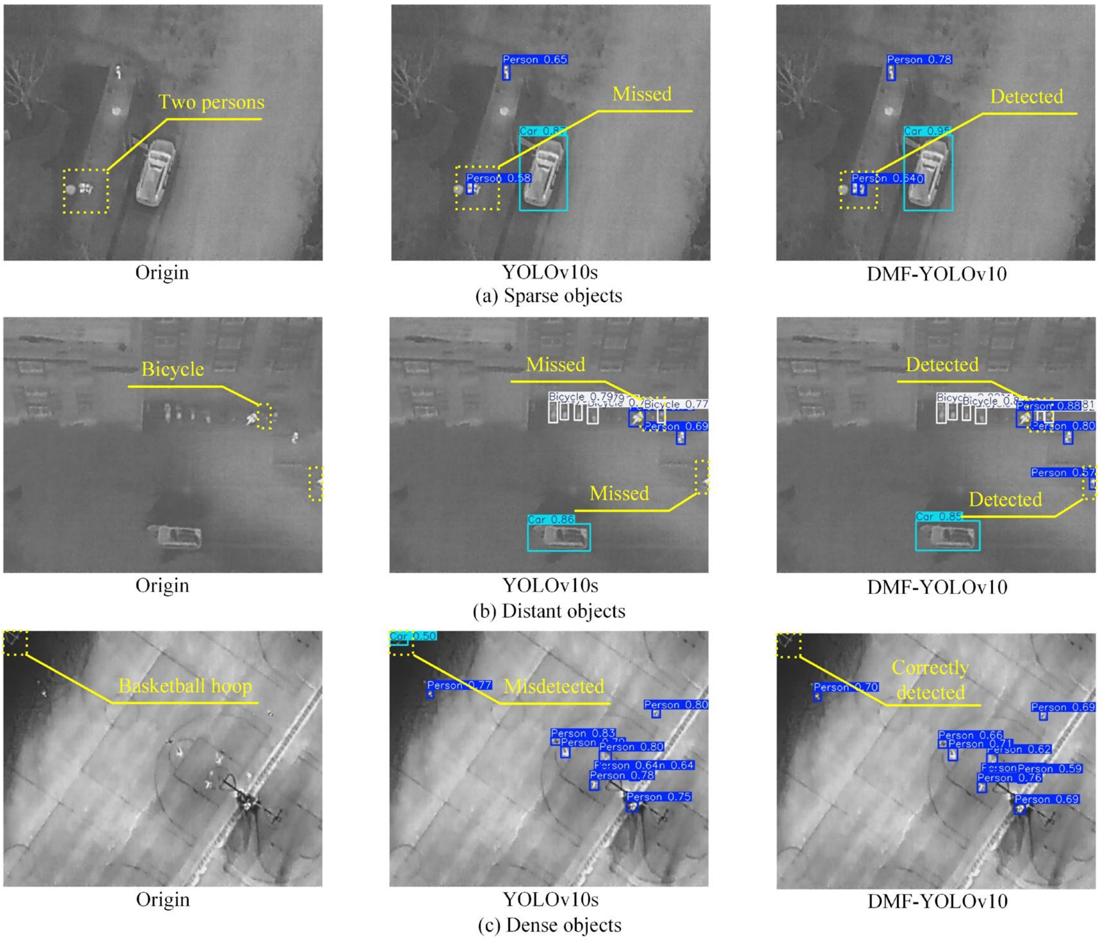


**Figure 17.** Comparison of detection effect in different scenes.

As shown in Table 6, the proposed DMF-YOLOv10 achieves superior performance compared to the baseline YOLOv10s model on the HIT-UAV dataset. Specifically, the mAP50 metric improves from 79.3% to 81.4%, and mAP50:95 increases from 49.5% to 52.8%. Although the parameter count rises from 12.5M to 14.2M and the inference speed increases from 3.8ms to 4.5ms, the detection accuracy is significantly enhanced. Compared to YOLOv10l and CA-YOLOv10, the proposed method reduces parameter counts by 39.57% and 67.13%, decreases inference time by 49.43% and 66.91%, respectively, and improves recognition accuracy for infrared small targets.

Figure 16 compares the detection performance of different models across five target categories (vertical axis: detection accuracy; horizontal axis: target categories). The results demonstrate that DMF-YOLOv10, benefiting from its dynamic multi-scale feature fusion mechanism tailored for complex scenarios, achieves optimal performance in most categories. Notably, it attains 97.9% and 97.4% accuracy for Car and Bicycle categories, significantly outperforming other models. While YOLOv10l, a larger-scale model, slightly underperforms DMF-YOLOv10 in Person and Other Vehicle categories, its 96.9% accuracy for Car still surpasses baseline models like YOLOv5s (91.5%) and YOLOv8s (91.8%), indicating that increased model depth positively impacts specific target detection. Notably, CA-YOLOv10 outperforms YOLOv10l in Bicycle detection by integrating a coordinate attention mechanism, highlighting the efficacy of attention mechanisms for small target detection.

Figure 17 visualizes detection results, confirming that DMF-YOLOv10 significantly improves recognition accuracy for small-scale targets (e.g., pedestrians, bicycles) compared to YOLOv10s. In Figure 17(b), DMF-YOLOv10 successfully detects a pedestrian at the right edge of the image, while YOLOv10s in Figure 17(c) falsely identifies a basketball hoop in the upper-left corner as a car. These results validate the algorithm’s robustness in complex scenarios and its strong adaptability to cross-domain datasets.

## 5. Conclusions

This paper proposes DMF-YOLO, an improved algorithm based on the YOLOv10 framework, to address the challenges of small target detection in UAV aerial images. By introducing Dynamic Dilated Snake Convolution (DDSConv), a Multi-scale Feature Aggregation Module (MFAM), and an Expanded Window-based Bounding Box Regression Loss Function (EW-BBRLF), the model significantly enhances detection capabilities for multi-scale targets, particularly micro-objects. Experimental results demonstrate that DMF-YOLO achieves 50.1% and 81.4% mAP50 on the VisDrone and HIT-UAV datasets, respectively, surpassing the baseline YOLOv10s by 27.1% and 2.6%, while increasing parameters by only 24.4% and 11.9%, validating the algorithm's balanced advantage between accuracy and efficiency. Visualization analyses further confirm the model's enhanced robustness in dense scenes, complex backgrounds, and long-distance scenarios, with notable improvements in small target feature extraction and localization precision.

Although DMF-YOLO achieves a favorable trade-off between parameter count and detection speed, computational overhead from dynamic convolution and multi-scale fusion modules remains a challenge. Future work may explore model compression techniques such as knowledge distillation, channel pruning, or dynamic network architecture design to meet real-time processing requirements on UAV edge devices. Additionally, UAVs often employ multimodal sensors (e.g., visible-light and infrared). Future research could investigate detection frameworks integrating multispectral or thermal imaging data to improve target recognition under complex lighting and adverse weather conditions.

**Author Contributions:** Conceptualization, X.Y.; methodology, X.Y.; software, Y.L. and S.S.; validation, S.S.; resources, Q.H.; writing—original draft preparation, W.Y.; writing—review and editing, H.Z.; project administration, Q.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China (61640308) and Natural Science Foundation of Hubei Province (2023AFB900).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Bogle, B.M.; Rosamond, W.D.; Snyder, K.T.; Zègre-Hemsey, J.K. The Case for Drone-Assisted Emergency Response to Cardiac Arrest: An Optimized Statewide Deployment Approach. N. C. Med. J. **2019,** 80, 204–212. https://doi.org/10.18043/ncm.80.4.204.
2. Raoult, V.; Colefax, A.P.; Allan, B.M.; Cagnazzi, D.; Castelblanco-Martínez, N.; Ierodiaconou, D.; Johnston, D.W.; Landeo-Yauri, S.; Lyons, M.; Pirotta, V.; et al. Operational Protocols for the Use of Drones in Marine Animal Research. Drones **2020**, 4, 64. https://doi.org/10.3390/drones4040064.
3. Potter, B.; Valentino, G.; Yates, L.; Benzing, T.; Salman, A. Environmental Monitoring Using a Drone-Enabled Wireless Sensor Network. In Proceedings of the 2019 Systems and Information Engineering Design Symposium (SIEDS); **2019**; pp. 1–6.
4. Monteiro, J.G.; Jiménez, J.L.; Gizzi, F.; Přikryl, P.; Lefcheck, J.S.; Santos, R.S.; Canning-Clode, J. Novel Approach to Enhance Coastal Habitat and Biotope Mapping with Drone Aerial Imagery Analysis. Sci. Rep. **2021**, 11, 574. https://doi.org/10.1038/s41598-020-80612-7.

5. Kyrkou, C.; Theocharides, T. EmergencyNet: Efficient Aerial Image Classification for Drone-Based Emergency Monitoring Using Atrous Convolutional Feature Fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **2020**, 13, 1687–1699. https://doi.org/10.1109/JSTARS.2020.2969809.
6. Ouattara, T.A.; Sokeng, V.-C.J.; Zo-Bi, I.C.; Kouamé, K.F.; Grinand, C.; Vaudry, R. Detection of Forest Tree Losses in Côte d'Ivoire Using Drone Aerial Images. Drones **2022**, 6, 83. https://doi.org/10.3390/drones6040083.
7. Degollada, E.; Amigó, N.; O'Callaghan, S.A.; Varola, M.; Ruggero, K.; Tort, B. A Novel Technique for Photo-Identification of the Fin Whale, Balaenoptera Physalus, as Determined by Drone Aerial Images. Drones **2023**, 7, 220. https://doi.org/10.3390/drones7030220.
8. Chen, J.; Wang, G.; Luo, L.; Gong, W.; Cheng, Z. Building Area Estimation in Drone Aerial Images Based on Mask R-CNN. IEEE Geosci. Remote Sens. Lett. **2021**, 18, 891–894. https://doi.org/10.1109/LGRS.2020.2988326.
9. Hmidani, O.; Ismaili Alaoui, E.M. A Comprehensive Survey of the R-CNN Family for Object Detection. In Proceedings of the 2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet); **2022**; pp. 1–6.
10. Xu, J.; Ren, H.; Cai, S.; Zhang, X. An Improved Faster R-CNN Algorithm for Assisted Detection of Lung Nodules. Comput. Biol. Med. **2023**, 153, 106470. https://doi.org/10.1016/j.compbiomed.2022.106470.
11. Fu, X.; Wei, G.; Yuan, X.; Liang, Y.; Bo, Y. Efficient YOLOv7-Drone: An Enhanced Object Detection Approach for Drone Aerial Imagery. Drones **2023**, 7, 616. https://doi.org/10.3390/drones7100616.
12. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo Algorithm Developments. Procedia Comput. Sci. **2022**, 199, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135.
13. Chen, Z.; Guo, H.; Yang, J.; Jiao, H.; Feng, Z.; Chen, L.; Gao, T. Fast Vehicle Detection Algorithm in Traffic Scene Based on Improved SSD. Measurement **2022**, 201, 111655. https://doi.org/10.1016/j.measurement.2022.111655.
14. Zhao, X.; Xia, Y.; Zhang, W.; Zheng, C.; Zhang, Z. YOLO-ViT-Based Method for Unmanned Aerial Vehicle Infrared Vehicle Target Detection. Remote Sens. **2023**, 15, 3778. https://doi.org/10.3390/rs15153778.
15. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); **2017**; pp. 936–944.
16. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2018**; pp. 8759–8768.
17. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A Global-Local Self-Adaptive Network for Drone-View Object Detection. IEEE Trans. Image Process. **2021**, 30, 1556–1569. https://doi.org/10.1109/TIP.2020.3045637.
18. Cai, D.; Lu, Z.; Fan, X.; Ding, W.; Li, B. Improved YOLOv4-Tiny Target Detection Method Based on Adaptive Self-Order Piecewise Enhancement and Multiscale Feature Optimization. Appl. Sci. **2023**, 13, 8177. https://doi.org/10.3390/app13148177.
19. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); **2021**; pp. 2778–2788.
20. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the Computer Vision – ECCV 2018; Springer: Cham, Switzerland, **2018**; pp. 122–138.
21. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2018**; pp. 4510–4520.
22. Saif, A.F.M.S.; Prabuwono, A.S.; Mahayuddin, Z.R. Moment Feature Based Fast Feature Extraction Algorithm for Moving Object Detection Using Aerial Images. PLOS ONE **2015**, 10, e0126212. https://doi.org/10.1371/journal.pone.0126212.
23. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); **2017**; pp. 764–773.

24. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); **2023**; pp. 13435–13444.
25. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); **2023**; pp. 14408–14419.
26. Qi, Y.; He, Y.; Qi, X.; Zhang, Y.; Yang, G. Dynamic Snake Convolution Based on Topological Geometric Constraints for Tubular Structure Segmentation. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); **2023**; pp. 6047–6056.
27. Niu, Y.; Fan, S.; Cheng, X.; Yao, X.; Wang, Z.; Zhou, J. Road Crack Detection by Combining Dynamic Snake Convolution and Attention Mechanism. Appl. Sci. **2024**, 14, 8100. https://doi.org/10.3390/app14188100.
28. Chen, J.; Jin, W.; Liu, Y.; Huang, X.; Zhang, Y. Multi-Scale and Dynamic Snake Convolution-Based YOLOv9 for Steel Surface Defect Detection. J. Supercomput. **2025**, 81, 541. https://doi.org/10.1007/s11227-025-07036-w.
29. Wang, S.; Jiang, H.; Yang, J.; Ma, X.; Chen, J. AMFEF-DETR: An End-to-End Adaptive Multi-Scale Feature Extraction and Fusion Object Detection Network Based on UAV Aerial Images. Drones **2024**, 8, 523. https://doi.org/10.3390/drones8100523.
30. Zhang, J.; Xie, J.; Gong, W. Object Detection Algorithm with Dual-Modal Rectification Fusion Based on Self-Guided Attention. J. Comput. Eng. Appl. **2023**, 36, 793. https://doi.org/10.16451/j.cnki.issn1003-6059.202309003.
31. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); **2022**; pp. 13658–13667.
32. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); **2019**; pp. 6568–6577.
33. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Han, K.; Wang, Y. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. Adv. Neural Inf. Process. Syst. **2023**, 36, 51094–51112.
34. Yu, J.; Wu, T.; Zhang, X.; Zhang, W. An Efficient Lightweight SAR Ship Target Detection Network with Improved Regression Loss Function and Enhanced Feature Information Expression. Sensors **2022**, 22, 3447. https://doi.org/10.3390/s22093447.
35. Zhu, C.; He, Y.; Savvides, M. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); **2019**; pp. 840–849.
36. Chen, X.; Li, L.; Li, Z.; Liu, M.; Li, Q.; Qi, H.; Ma, D.; Wen, Y.; Cao, G.; Yu, P.L.H. KD Loss: Enhancing Discriminability of Features with Kernel Trick for Object Detection in VHR Remote Sensing Images. Eng. Appl. Artif. Intell. **2024**, 129, 107641. https://doi.org/10.1016/j.engappai.2023.107641.
37. Gevorgyan, Z. SIoU Loss: More Powerful Learning for Bounding Box Regression. arXiv **2022**, arXiv:2205.12740.
38. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); **2020**; pp. 9756–9765.
39. Ma, S.; Xu, Y. MPDIoU: A Loss for Efficient and Accurate Bounding Box Regression. arXiv **2023**, arXiv:2307.07662.
40. Zhang, H.; Xu, C.; Zhang, S. Inner-IoU: More Effective Intersection over Union Loss with Auxiliary Bounding Box. arXiv **2023**, arXiv:2311.02824.
41. Wang, M.; Liang, Z.; Huang, H.; Liang, A.; Sun, H.; Zhao, Y. Research and Application of YOLOv10 Algorithm Based on Image Recognition. In Proceedings of the 5th International Conference on Artificial Intelligence and Computer Engineering; ACM: New York, NY, USA, **2025**; pp. 535–540.

42. Guan, S.; Lin, Y.; Lin, G.; Su, P.; Huang, S.; Meng, X.; Liu, P.; Yan, J. Real-Time Detection and Counting of Wheat Spikes Based on Improved YOLOv10. Agronomy **2024**, 14, 1936. https://doi.org/10.3390/agronomy14091936.
43. Wang, Q.; Wang, X.; Hou, J.; Liu, X.; Wen, H.; Ji, Z. MF-YOLOv10: Research on the Improved YOLOv10 Intelligent Identification Algorithm for Goods. Sensors **2025**, 25, 2975. https://doi.org/10.3390/s25102975.
44. Samma, H.; Suandi, S.A.; Ismail, N.A.; Sulaiman, S.; Ping, L.L. Evolving Pre-Trained CNN Using Two-Layers Optimizer for Road Damage Detection from Drone Images. IEEE Access **2021**, 9, 158215–158226. https://doi.org/10.1109/ACCESS.2021.3131231.
45. Lee, D.-H. CNN-Based Single Object Detection and Tracking in Videos and Its Application to Drone Detection. Multimed. Tools Appl. **2021**, 80, 34237–34248. https://doi.org/10.1007/s11042-020-09924-0.
46. Wang, Z.; Dang, C.; Zhang, R.; Wang, L.; He, Y.; Wu, R. MDDFA-Net: Multi-Scale Dynamic Feature Extraction from Drone-Acquired Thermal Infrared Imagery. Drones **2025**, 9, 224. https://doi.org/10.3390/drones9030224.
47. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); October 2019; pp. 213–226.
48. Suo, J.; Wang, T.; Zhang, X.; Chen, H.; Zhou, W.; Shi, W. HIT-UAV: A High-Altitude Infrared Thermal Dataset for Unmanned Aerial Vehicle-Based Object Detection. Sci Data 2023, 10, 227. https://doi.org/10.1038/s41597-023-02066-6.
49. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); **2017**; pp. 2999–3007.
50. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; **2018**; pp. 6154–6162.
51. Sun, C.; Zhang, S.; Qu, P.; Wu, X.; Feng, P.; Tao, Z.; Zhang, J.; Wang, Y. MCA-YOLOV5-Light: A Faster, Stronger and Lighter Algorithm for Helmet-Wearing Detection. Appl. Sci. **2022**, 12, 9697. https://doi.org/10.3390/app12199697.
52. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. DAMO-YOLO: A Report on Real-Time Object Detection Design. arXiv **2023**, arXiv:2301.13566.
53. Shen, L.; Lang, B.; Song, Z. CA-YOLO: Model Optimization for Remote Sensing Image Object Detection. IEEE Access **2023**, 11, 64769–64781. https://doi.org/10.1109/ACCESS.2023.3290481.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.