remote sensing

MDPI

Article

YOLO-SRMX: A Lightweight Model for Real-Time Object Detection on Unmanned Aerial Vehicles

Shimin Weng [1], Han Wang [1], Jiashu Wang [1], Changming Xu [1,*] and Ende Zhang [2]

[1] School of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066004, China; 202216286@stu.neu.edu.cn (S.W.); 202212198@stu.neu.edu.cn (H.W.); 202212719@stu.neu.edu.cn (J.W.)
[2] School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China; zed@cc.neu.edu.cn
* Correspondence: changmingxu@neuq.edu.cn

Abstract

Unmanned Aerial Vehicles (UAVs) face a significant challenge in balancing high accuracy and high efficiency when performing real-time object detection tasks, especially amidst intricate backgrounds, diverse target scales, and stringent onboard computational resource constraints. To tackle these difficulties, this study introduces YOLO-SRMX, a lightweight real-time object detection framework specifically designed for infrared imagery captured by UAVs. Firstly, the model utilizes ShuffleNetV2 as an efficient lightweight backbone and integrates the novel Multi-Scale Dilated Attention (MSDA) module. This strategy not only facilitates a substantial 46.4% reduction in parameter volume but also, through the flexible adaptation of receptive fields, boosts the model's robustness and precision in multi-scale object recognition tasks. Secondly, within the neck network, multi-scale feature extraction is facilitated through the design of novel composite convolutions, ConvX and MConv, based on a "split–differentiate–concatenate" paradigm. Furthermore, the lightweight GhostConv is incorporated to reduce model complexity. By synthesizing these principles, a novel composite receptive field lightweight convolution, DRFAConvP, is proposed to further optimize multi-scale feature fusion efficiency and promote model lightweighting. Finally, the Wise-IoU loss function is adopted to replace the traditional bounding box loss. This is coupled with a dynamic non-monotonic focusing mechanism formulated using the concept of outlier degrees. This mechanism intelligently assigns elevated gradient weights to anchor boxes of moderate quality by assessing their relative outlier degree, while concurrently diminishing the gradient contributions from both high-quality and low-quality anchor boxes. Consequently, this approach enhances the model's localization accuracy for small targets in complex scenes. Experimental evaluations on the HIT-UAV dataset corroborate that YOLO-SRMX achieves an mAP_{50} of 82.8%, representing a 7.81% improvement over the baseline YOLOv8s model; an F1 score of 80%, marking a 3.9% increase; and a substantial 65.3% reduction in computational cost (GFLOPs). YOLO-SRMX demonstrates an exceptional trade-off between detection accuracy and operational efficiency, thereby underscoring its considerable potential for efficient and precise object detection on resource-constrained UAV platforms.

Keywords: Unmanned Aerial Vehicle (UAV); object detection; lightweight model; YOLO; real-time detection

check for updates

Academic Editors: Martin Gade, Yi Wan, Wensheng Wang and Zhirui Wang

Received: 22 May 2025
Revised: 24 June 2025
Accepted: 3 July 2025
Published: 5 July 2025

Citation: Weng, S.; Wang, H.; Wang, J.; Xu, C.; Zhang, E. YOLO-SRMX: A Lightweight Model for Real-Time Object Detection on Unmanned Aerial Vehicles. *Remote Sens.* **2025**, *17*, 2313. https://doi.org/10.3390/rs17132313

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Over the past few years, Unmanned Aerial Vehicle (UAV) technology has undergone rapid and significant development. Due to their exceptional flexibility, stealth, and high efficiency, UAVs have become indispensable reconnaissance and detection tools in both civilian and military domains, with a broad spectrum of applications. UAV-based object detection technologies have been widely applied across diverse sectors, including agriculture, surveying and mapping, disaster management, healthcare, firefighting, and national defense. As illustrated in Figure 1, researchers have extensively studied UAV object detection applications in these fields. For example, in agriculture, F. Ahmad et al. [1] performed field studies to investigate how UAV sprayer configurations influence spray deposition patterns in both target and non-target zones for weed control purposes. Their findings demonstrated how UAVs enabled the optimization of spraying efficacy and the reduction of pesticide drift. In surveying and mapping, M. Gašparović et al. [2] utilized UAVs equipped with low-cost RGB cameras to acquire high-resolution imagery and combined photogrammetric processing with machine-learning classification techniques to develop an automated method for generating high-precision weed distribution maps in farmland, providing crucial spatial information support for precision agriculture. In disaster management, N. D. Nath et al. [3] employed UAV video for spatial mapping. By integrating SIFT feature matching, progressive homography estimation, multi-object tracking, and iterative closest point registration, they achieved GPS-independent, high-precision localization and mapping of features in disaster-stricken areas. In firefighting, Kinaneva et al. [4] proposed an early forest fire detection method combining UAV platforms with artificial-intelligence techniques. Their approach utilized thermal imaging and real-time image analysis to enhance the identification of initial fire indicators, thereby improving the timeliness and accuracy of fire response in forest fire prevention, emergency response, and disaster management. In national defense, He et al. [5] introduced the YOLOv5s-pp algorithm, which integrates channel attention (CA) modules to optimize small-object detection from UAV perspectives, thereby boosting the robustness and safety of UAV platforms during essential operations, including military reconnaissance, security monitoring, and route patrolling.

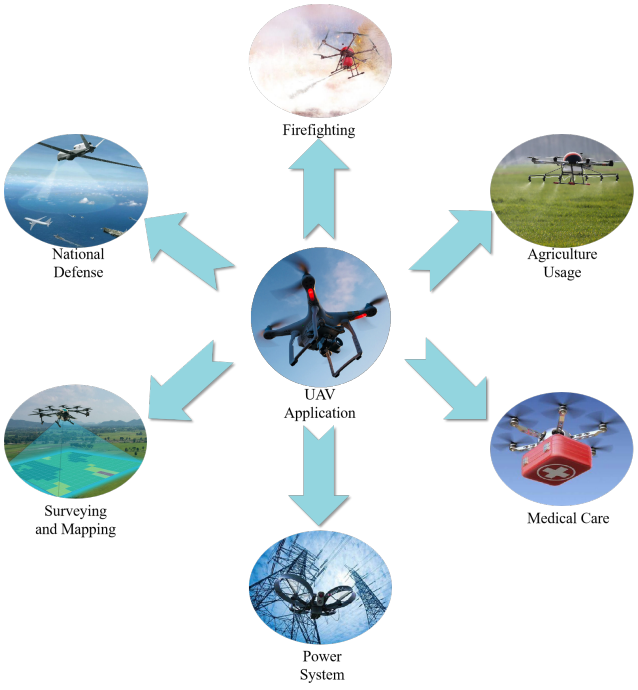


Figure 1. Drone Application Fields.

Currently, Convolutional Neural Networks (CNNs) stand as the predominant framework in the domain of object detection. CNN-based object detection methods can generally be categorized into two-stage and single-stage approaches. Two-stage algorithms, such as R-CNN [6], SPPNet [7], Fast R-CNN [8], and Faster R-CNN [9], depend on region proposal generation and generally achieve higher localization accuracy, despite being hindered by relatively slow inference speeds. Single-stage algorithms, such as SSD [10], YOLO [11] and RetinaNet [12], directly regress object locations and categories, achieving faster detection speeds. Their detection performance, especially for small objects, has been continuously improved through techniques like multi-scale feature fusion. Owing to their excellent trade-off between being lightweight and highly accurate, YOLO series algorithms have found widespread application in UAV object detection, spurring further research in this domain (e.g., YBUT [13]).

However, object detection from a UAV perspective typically faces several challenges: (1) complex and cluttered backgrounds; (2) highly variable object scales; (3) large fields of view (FOV) that yield sparse or uneven object distributions; (4) frequent changes in object location, orientation, and angle; and (5) stringent limits on onboard computational resources. These factors significantly impede the performance of UAV-based detectors. To address these challenges, researchers have introduced numerous methodologies adapted for distinct UAV operational environments. To mitigate background complexity in UAV imagery, Bo et al. [14] employed an enhanced YOLOv7-tiny network architecture. By redesigning the anchor box mechanism and incorporating key modules such as SPPFCSPC-SR, InceptionNeXt, and Get-and-Send, they improved feature representation for small targets and cross-level information fusion. This method significantly boosted detection precision and system resilience in complex dynamic environments, though its feature extraction capability remains limited in highly challenging scenarios. Similarly targeting background complexity, Wang et al. [15] proposed OATF-YOLO, which introduces an Orthogonal Channel Attention mechanism (OCAC2f), integrates a triple feature encoder and a scale-sequence feature fusion module (TESSF), and incorporates an internal factor (I-MPDIoU) into the loss function to enhance feature extraction and background discrimination. However, its robustness remains insufficient for aerial images with variable viewpoints, arbitrary orientations, complex illumination, and diverse weather conditions.

To cope with small, low-resolution targets, Zhao et al. [16] presented Subtle-YOLOv8. Building on YOLOv8, it integrates an Efficient Multi-Scale Attention module (EMA), Dynamic Serpentine Convolution (DSConv), a dedicated small-object detection head, and a Wise-IoU loss function. Although it markedly improves small-object detection, its performance degrades for slightly larger targets (e.g., trucks), and the overall model complexity increases. Similarly, Chen et al. [17] developed a Residual Progressive Feature Pyramid Network (R-AFPN). It fuses features via a Residual Progressive Feature Fusion (RAFF) module for adaptive spatial and cross-scale connections, employs a Shallow Information Extraction (SIE) module to capture fine low-level details, and uses a Hierarchical Feature Fusion (HFF) module to enrich deep features through bottom-up incremental fusion. While R-AFPN excels at single-class, small-object detection with low parameter counts, its robustness on multi-class, large-scale-variation datasets (e.g., VisDrone2019) remains inferior to some state-of-the-art networks. To correct severe radial distortion in wide-angle and fisheye images, Kakani et al. [18] devised an outlier-optimized automatic distortion-correction method. By extracting and aggregating straight-line segments into line-member sets, introducing an iterative outlier-optimization mechanism, and applying an accumulated vertical-line-angle loss, they jointly optimize distortion parameters and purify line members. Although this significantly improves corrected image quality and downstream vision tasks, the method's robustness degrades in scenes lacking abundant linear features. To fulfill timeliness stipulations amidst constrained resources, Ma et al. [19] proposed LW-YOLOv8, a lightweight detector based on an improved YOLOv8s. They designed a CSP-CTFN module combining CNNs with multi-head self-attention and a parameter-sharing

PSC-Head detection head, and adopted SIoU loss to boost localization precision. While this approach drastically reduces model size and accelerates inference, its adaptability on extremely resource-limited devices and in complex environments still needs enhancement. Within the YOLOv5 framework, Chen et al. [20] replaced the backbone's C3 modules with Efficient Layer Shuffle Aggregation Networks (ELSANs) to improve speed and efficiency, and integrated Partial Convolution (PCConv) with an SE attention mechanism in the neck to further reduce computation and increase accuracy. This method significantly improves detection in heavily overlapping and occluded agricultural scenes, but its performance declines under extreme overlap, and false positives persist in complex backgrounds. Finally, the issue of object rotation in UAV imagery has also been addressed. Pan et al. [21] extended YOLOv8 by adding a rotation-angle (θ) output dimension in the detection head to remedy object deformation and rotation, as seen from UAV viewpoints. Li et al. [22] improved the S2A-Net by embedding PSA/CIM modules for richer feature representation and combining D4 dehazing with an image-weighted sampling strategy for optimized training. Although this enhances rotated-object detection, its generalization ability and runtime efficiency remain areas in need of further optimization.

Despite substantial advancements in the domain of UAV object detection, investigations that effectively integrate multi-scale feature fusion strategies with compact network designs to synergistically enhance both detection accuracy and operational efficiency remain relatively limited. To tackle these challenges and achieve a better balance between high-precision detection and model lightweighting, this work introduces a specialized lightweight UAV object detection approach. The main contributions of this paper are as follows:

In response to the resource constraints and real-time requirements of UAV platforms, this paper constructs a systematic lightweight detection architecture. Firstly, an efficient ShuffleNetV2 is adopted as the backbone network to significantly reduce the base parameter count and computational costs. To compensate for the potential decline in feature representation capability caused by deep lightweight design, particularly in capturing contextual information of small targets against complex backgrounds, a Multi-Scale Dilated Attention (MSDA) mechanism is further integrated into the backbone network. This mechanism effectively enhances the feature extraction efficiency of the backbone network through its dynamic multi-scale receptive field capability. The synergistic optimization of this backbone and the lightweight design of the neck network result in a remarkable 65.3% reduction in the model's computational load (GFLOPs) compared to the baseline model.

To address the challenge of high false detection rates caused by the low resolution and limited feature information of small targets, this paper abandons the conventional approach of stacking or replacing modules. Instead, it proposes a novel neck network design paradigm centered on the concept of "heterogeneous feature decoupling and dynamic aggregation". This paradigm aims to maximize the model's ability to extract and fuse diverse features at minimal computational cost, achieved through our original two core components:

- Original bottleneck modules for heterogeneous decoupling (C2f_ConvX & C2f_MConv): This paper introduces a novel "split-differentiate-merge" strategy within the C2f architecture. Unlike standard group convolution's homogeneous processing, the proposed C2f_ConvX and C2f_MConv modules asymmetrically split the information flow and apply convolutional transformations with divergent scales and complexities. This asymmetric design decouples and extracts richer gradient combinations and feature spectra from raw features, significantly enhancing the discernibility of small targets in complex backgrounds without a substantial parameter increase.
- Original lightweight convolution for dynamic aggregation (DRFAConvP): To efficiently fuse the aforementioned decoupled heterogeneous features, this paper proposes DRFAConvP—a fundamental advancement over existing lightweight convolutions (e.g.,

GhostConv). While GhostConv relies on static linear transformations for feature generation, DRFAConvP innovatively integrates an input-dependent dynamic attention mechanism into this process. This enables adaptive aggregation of multi-receptive-field information, where the network dynamically focuses on the most discriminative spatial details for the current target, effectively resolving the severe scale variation challenges in UAV-based detection.

To enhance the model's small-target localization precision in complex scenarios, the Wise-IoU (WIoU) loss function is introduced, incorporating a dynamic non-monotonic focusing mechanism. This mechanism intelligently modulates gradient weights according to anchor boxes' outlier degrees, prioritizing the optimization of medium-quality anchors while suppressing harmful gradients from low-quality samples, thereby achieving more accurate small-target localization in cluttered backgrounds and effectively reducing both missed detection and false alarm rates.

2. Model Design

2.1. YOLOv8 Architecture Analysis

YOLOv8 is a family of YOLO models released by Ultralytics in 2023. It can be applied to multiple computer vision tasks, including image classification, object detection, instance segmentation, and keypoint detection. The model is available in five different sizes, ranging from the compact YOLOv8n to the large YOLOv8x, with detection accuracy increasing alongside model size. The YOLOv8 architecture consists of four primary components: Input, Backbone, Neck, and Head. The Input module preprocesses images for efficient network processing, employing adaptive scaling to improve both image quality and inference speed [23]. The Backbone consists of Conv, C2f, and SPPF modules that collaboratively extract features from the input. Notably, the C2f module integrates two convolutional layers and multiple Bottleneck blocks to capture richer gradient information [24], while the SPPF module applies three successive max-pooling operations to harvest multi-scale object features, thereby enhancing detection accuracy across various object sizes [25]. The overall YOLOv8 architecture is illustrated in Figure 2.

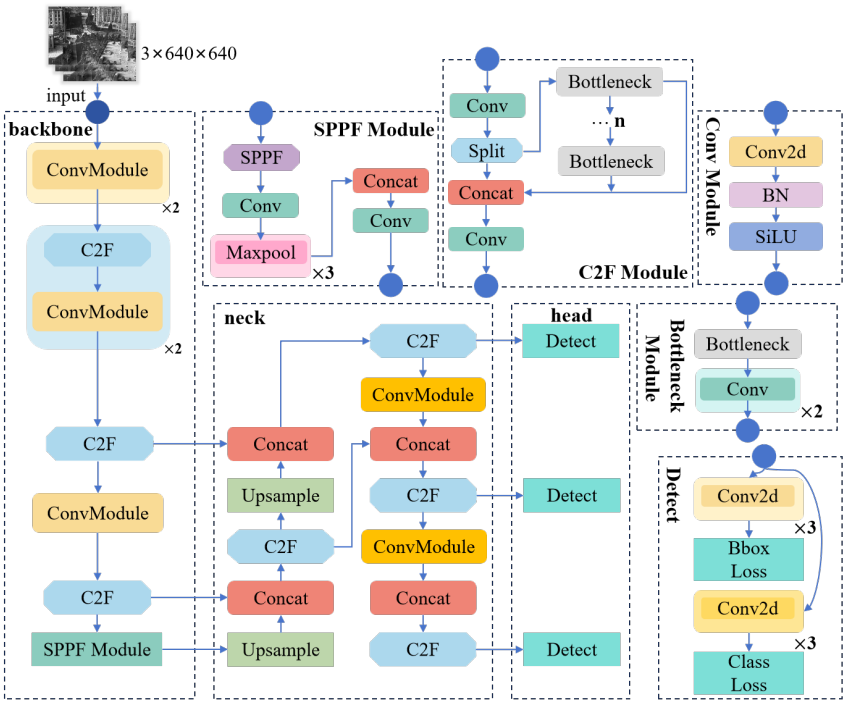


Figure 2. YOLOv8 architecture diagram.

2.2. Improved YOLOv8s Model Network Structure

Object detection tasks performed from a UAV perspective face multiple inherent challenges. Firstly, high-altitude, long-distance imaging and a wide field of view (FoV) often result in targets appearing small, with low resolution. Their faint features can easily be confounded with complex backgrounds containing diverse terrains (such as buildings, vegetation, and textures), severely hindering accurate identification and localization. Secondly, the inherent limitations in computational power, storage space, and energy consumption of UAV platforms impose stringent demands on the lightweight design and computational efficiency of detection models. Therefore, while pursuing high detection accuracy, it is crucial to consider the resource efficiency for model deployment; achieving an effective balance between these two aspects is key in this research field.

To surmount these difficulties, we present YOLO-SRMX, an efficient detection architecture customized for infrared targets, built upon YOLOv8s. Figure 3 delineates the refined YOLOv8 structure, which comprises the following augmentations:

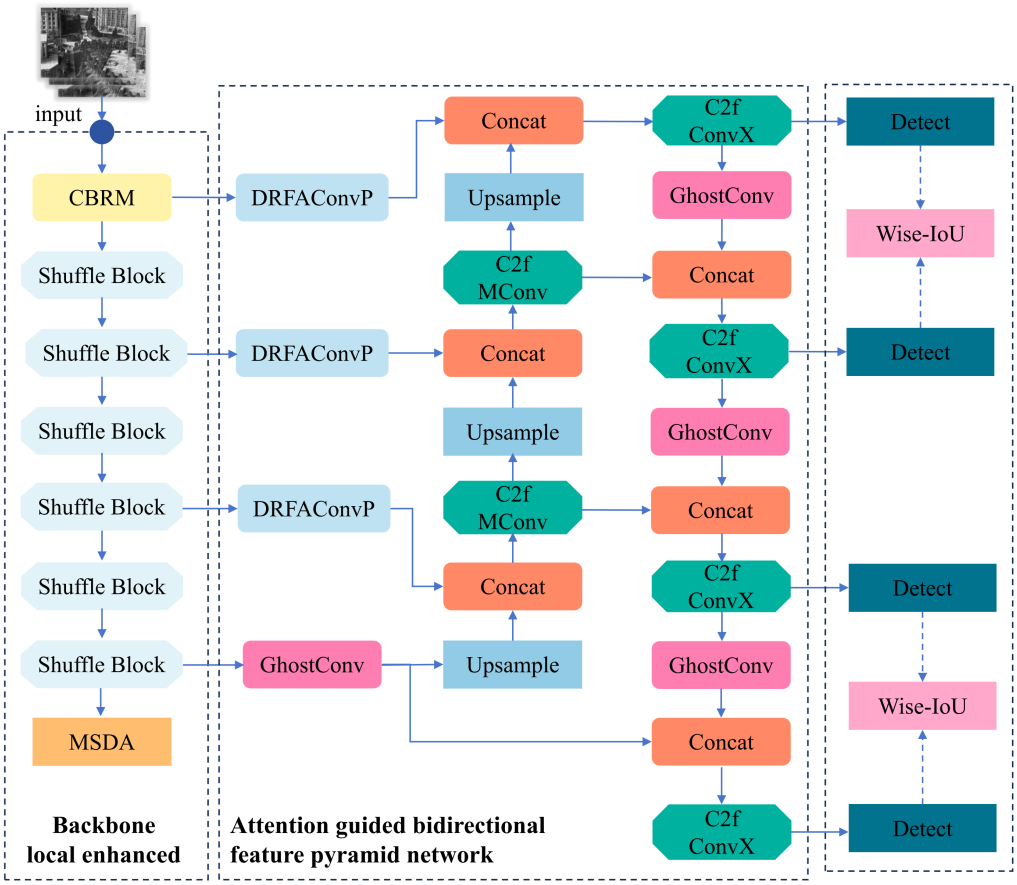


Figure 3. Improved YOLOv8 architecture.

- In the Backbone network, an efficient ShuffleNetV2 architecture is employed and integrated with the MSDA mechanism. This design significantly reduces computational overhead while leveraging MSDA's dynamic focusing and contextual awareness capabilities to strengthen the extraction efficacy of critical features.
- In the Neck network, several optimizations are implemented to enhance feature fusion and reduce computational load. Novel composite convolutions, ConvX and MConv, along with their corresponding C2f_ConvX and C2f_MConv modules, are designed based on a "split–differentiate–process–concatenate" strategy. Additionally, the lightweight GhostConv is introduced, and by combining the multi-branch pro-

cessing principles of composite convolutions with the lightweight characteristics of GhostConv, a novel Dynamic Receptive Field Attention Convolution, DRFAConvP, is proposed. These collective improvements optimize the neck network's computational efficiency and feature processing capabilities.

- In the Loss function component, the original loss function is replaced with Wise-IoU (WIoU). It incorporates a dynamic non-monotonic focusing approach. Through this approach, gradient weights are judiciously apportioned based on the respective outlier degree of each anchor box. The optimization of anchor boxes with intermediate quality is emphasized, whereas the gradient influence from samples of poor quality is mitigated. This allows the model to achieve more exact localization of multi-scale targets and bolsters its overall stability.

2.3. Lightweight Feature Extraction Network Fusing Attention Mechanism

In practical application deployment, especially on mobile platforms with limited computational power and memory resources, such as unmanned aerial vehicles (UAVs), model deployment faces significant challenges. Large-scale neural network algorithms are often difficult to run effectively in these environments due to their high resource demands. Therefore, optimizing model parameter count and computational complexity is a key prerequisite for ensuring deployment feasibility and operational efficiency in resource-constrained scenarios.

Although the original YOLOv8 model exhibits good detection performance, it has inherent defects in losing key target information during the feature extraction stage. Coupled with its large model size and high computational requirements, its effective deployment on resource-constrained devices (such as UAVs) is limited. To overcome these limitations, this paper makes critical improvements to the backbone network of YOLOv8s: it adopts the lightweight and computationally efficient ShuffleNetV2 as the basic architecture, leveraging its unique Channel Shuffle and Group Convolution mechanisms to significantly reduce model complexity and parameter count. Concurrently, the MSDA mechanism is integrated to enhance the network's dynamic capture and perception capabilities for multi-scale key features. This alternative backbone solution, combining ShuffleNetV2 and MSDA, aims to achieve a significant reduction in parameter count and computational load and improve the effectiveness of feature extraction, thereby constructing a lightweight and efficient detection model more suitable for resource-constrained application scenarios.

2.3.1. ShuffleNetV2 Module

Ma et al. [26] proposed a set of design guidelines for optimizing efficiency metrics, addressing the issue of existing networks being overly reliant on indirect indicators like FLOPs while neglecting actual operational speed. Based on these guidelines, They introduced the ShuffleNetV2 network module, evolving from their prior work on the ShuffleNetV1 model. Figure 4 shows that this module fundamentally comprises two main sections: a basic unit and a downsampling unit.

For the basic unit, as shown in Figure 4a, input channels C are separated into dual pathways (typically $C/2$ and $C/2$). One path serves as an identity mapping, which is directly passed through, effectively achieving feature reuse. The other path processes features using a combination of a standard convolution (kernel size 1×1), a depthwise separable convolution (kernel size 3×3), and activation layers. This design enables efficient cross-channel feature fusion while minimizing computational overhead. To prevent network fragmentation and improve operational efficiency, feature channels are not evenly divided. Furthermore, a serial structure is used to ensure consistent channel widths across

all convolutional layers, thereby minimizing memory access costs. Channel shuffle is also employed to facilitate effective communication between different channel groups.

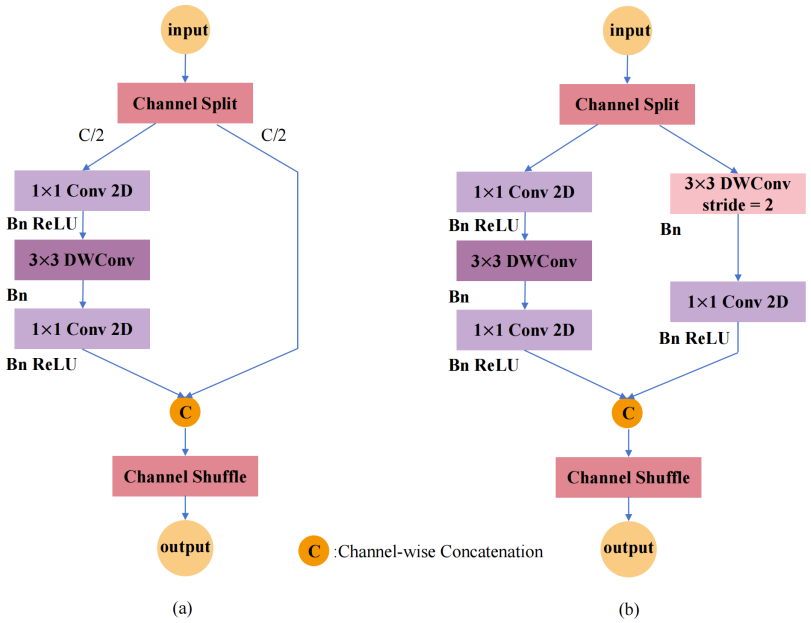


Figure 4. ShuffleNetV2 Unit. (**a**) Basic unit. (**b**) Downsampling unit.

The downsampling unit, shown in Figure 4b, differs from the basic unit by omitting the channel split operation, meaning all input c channels participate in computation. Concurrently, the branch that performs identity mapping in the basic unit introduces a depthwise separable convolution (kernel size 3×3), a standard convolution (kernel size 1×1), and activation layers. The downsampling unit reduces the spatial dimensions of the feature map by half, concurrently doubling the number of channels. Subsequent to combining the two branches, a channel rearrangement technique is employed to facilitate an exchange of data across channels. By directly increasing the network's width and the number of feature channels, the downsampling unit enhances feature extraction capabilities without significantly increasing computational costs.

2.3.2. MSDA Attention Mechanism

Object detection tasks in remote sensing images face two inherent challenges: first, complex and variable backgrounds can easily interfere with targets; second, target scales vary greatly, ranging from small vehicles to large areas. Traditional convolutional operations, owing to their fixed receptive field sizes, struggle to efficiently acquire both the fine-grained attributes of diminutive objects alongside the holistic structural data of macroscopic objects in parallel. Although the aforementioned ShuffleNetV2 backbone network significantly improves the model's computational efficiency, laying a foundation for lightweight deployment, its standard structure still has limitations in adequately addressing the feature representation challenges posed by extreme scale variations and strong background interference in remote sensing images. To precisely address these obstacles and subsequently bolster the robustness and distinctiveness of feature extraction, this study judiciously incorporates the MSDA module subsequent to the introduction of ShuffleNetV2.

The core idea of MSDA is to divide the channel dimension of the input feature map into n groups (or "heads") and apply a Scaled Dilated Attention (SWDA) operation with different dilation rates within each group, as shown in Figure 5. This multi-head, multi-dilation rate design enables the model to

- Parallelly capture multi-scale contextual information: heads with a dilation rate of 1 focus on local, nearby details, while heads with larger dilation rates (e.g., r = 3, 5) can capture longer-range sparse dependencies, effectively expanding the receptive field.
- Dynamically adjust attention regions: through a self-attention mechanism, the importance weights of features within each head are calculated, assigning greater weights to more relevant features (such as target regions) while suppressing interference from irrelevant background information.

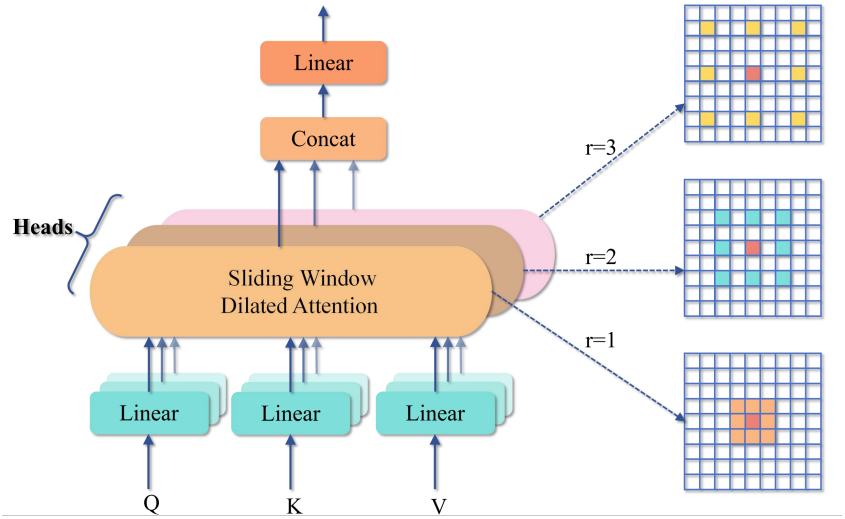


Figure 5. Overall Architecture of MSDA.

The SWDA operation applied to each attention head i can be represented by the following equation:

$$H_a = \text{SWDA}(Q_a, K_a, V_a, r_a), \quad 1 \le a \le n \tag{1}$$

where Q_a, K_a, V_a denote the query, key, and value in order. The dilation rate specific to each head is designated by r_a (e.g., $r_1 = 1, r_2 = 3$). The output generated by each attention head is identified as H_a.

For a position (a, b) in the original feature map, before the feature map is split, the output component X_{ab} from the SWDA operation is

$$X_{ab} = \text{Attention}(Q_{ab}, K_r, V_r) = \text{Softmax}\left(\frac{Q_{ab}K_r^T}{\sqrt{d_k}}\right) \times V_r, \quad 1 \le a \le w, 1 \le b \le w \tag{2}$$

where the feature map's height and width are denoted by H and w, respectively. K_r and V_r comprise the key and value aggregates derived from feature maps K and V, governed by the dilation rate r. When considering a query Q_{ab} at coordinates (a, b), the key-value pairs involved in its attention calculation originate from the coordinate set (a', b'), detailed as follows:

$$\{(a', b') \mid a' = a + p \times r, b' = b + q \times r\}, -\frac{w}{2} \le p, q \le \frac{w}{2}, p, q \in \mathbb{Z}^+ \tag{3}$$

where w denotes the size of the local window (or kernel) employed for key and value selection, while r specifies the dilation rate of this attention head.

After extracting multi-scale features through SWDA operations with different dilation rates, the MSDA module concatenates all head outputs H_a and performs effective feature fusion through a linear layer. This procedure enables the model to synthesize data from varied receptive fields, thereby encompassing both fine-grained local specifics and comprehensive global architectures.

Through the parallel deployment of attention heads with diverse dilation rates, the MSDA mechanism empowers the network to concurrently acquire contextual data spanning various extents. This significantly bolsters the discernment of fine points in diminutive targets and the comprehension of broad structures in large targets. This adaptable receptive field significantly enhances the network's capacity to accommodate substantial variations in target dimensions within remote sensing imagery [27]. Furthermore, the built-in attention mechanism of MSDA further optimizes the features extracted by ShuffleNetV2. By adaptively assigning higher weights to key features and suppressing background noise, it significantly improves the network's ability to distinguish targets in cluttered backgrounds.

To objectively assess the practical effectiveness of the MSDA mechanism within our framework and to clarify its performance differences relative to current mainstream attention techniques, we conducted a systematic comparison—under a unified experimental benchmark—between MSDA and representative modules such as Global Attention Mechanism (GAM) [28], inverted Residual Mobile Block (iRMB) [29], Coordinate Attention Moudule (CA) [30], Efficient Channel Attention (ECA) [31], and Convolutional Block Attention Moudule (CBAM) [32] . The specific results of this comparison are presented and analyzed in the Model Training and Evaluation section.

The synergy between ShuffleNetV2's efficiency and MSDA's multi-scale perceptual and attention-focusing strengths collectively forms the novel lightweight and highly expressive backbone architecture proposed in this paper, thereby substantially boosting the model's object detection accuracy and robustness in intricate conditions.

2.4. Improved Feature Fusion Network Design

To refine the YOLOv8 neck component (Neck) to meet the demands of high efficiency and robust feature fusion in UAV applications, this study presents two principal enhancements: firstly, the integration of lightweight convolutions, which curtail the complexity of fundamental computational units by utilizing more streamlined feature extraction and rearrangement mechanisms; secondly, the development of innovative multi-scale feature fusion modules intended to supersede the conventional C2f structure, thereby bolstering the acquisition and integration capacities for multi-scale data through the optimization of feature extraction and aggregation strategies within parallel branches.

2.4.1. Lightweight Convolution: GhostConv Module

In the YOLOv8 network architecture, the feature fusion module employs a Cross-Stage Partial (CSP) network. Although CSP modules can reduce the parameter count to some extent, for efficient deployment of object detection tasks on UAV platforms with limited computational resources, the overall scale of the YOLOv8 network remains excessively large.

In the current research context of neural network model compression and pruning, extensive research has shown that feature maps generated by many mainstream convolutional neural network architectures often contain rich redundant information, manifesting as highly similar content in some features. Addressing this common phenomenon, GhostNet [33] proposes a viewpoint: these seemingly redundant features are crucial for ensuring certain neural networks achieve high-performance detection, and these features can be generated more efficiently. To achieve this goal, GhostNet [33] introduces the GhostConv module, as shown in Figure 6.

The fundamental concept behind the Ghost module involves breaking down the standard convolution operation into a two-stage process: First, it processes the input feature map $X \in \mathbb{R}^{c \times h \times w}$ using a standard (or backbone) convolutional layer to generate a few basic, representative intrinsic feature maps $Y' \in \mathbb{R}^{n \times h' \times w'}$. Next, the Ghost module applies a series of inexpensive linear transformations (such as depthwise separable convolutions

or simple pointwise operations). Each intrinsic feature map undergoes $l-1$ linear transformations and one identity mapping, thereby deriving additional ghost feature maps.The formation of the ghost feature map Y' is articulated by Equation (4):

$$Y' = \Phi_{a,b}(y'_a), \quad \forall a = 1, \ldots, n, \quad b = 1, \ldots, l \tag{4}$$

As per Equation (4), n indicates the number of channels in Y', with y'_a representing the a-th channel, and b referring to the b-th linear transformation applied to y'_a. For each feature map y'_a, $l-1$ linear transformations are executed, followed by a single identity transformation, thereby yielding an output Y with $m = n \cdot l$ channels. Hence, the total number of linear transformations performed on the feature maps is $n \cdot (l-1)$. Assuming a kernel dimension of $d \times d$, Equations (5) and (6) detail the Ghost module's theoretical computational speedup ratio r_s and its parameter compression ratio r_c:

$$r_s = \frac{m \cdot h' \cdot w' \cdot c \cdot k^2}{\frac{m}{l} \cdot h' \cdot w' \cdot c \cdot k^2 + (l-1)\frac{m}{l} \cdot h' \cdot w' \cdot d^2} \approx \frac{l \cdot c \cdot k^2}{c \cdot k^2 + (l-1)d^2} \approx l \tag{5}$$

$$r_c = \frac{m \cdot c \cdot k^2}{\frac{m}{l} \cdot c \cdot k^2 + (l-1)\frac{m}{l} \cdot d^2} \approx \frac{l \cdot c \cdot k^2}{c \cdot k^2 + (l-1)d^2} \approx l \tag{6}$$

From the above equations, it can be seen that Ghost convolution reduces the number of parameters by approximately l times without significantly impairing the network's ability to effectively extract features.

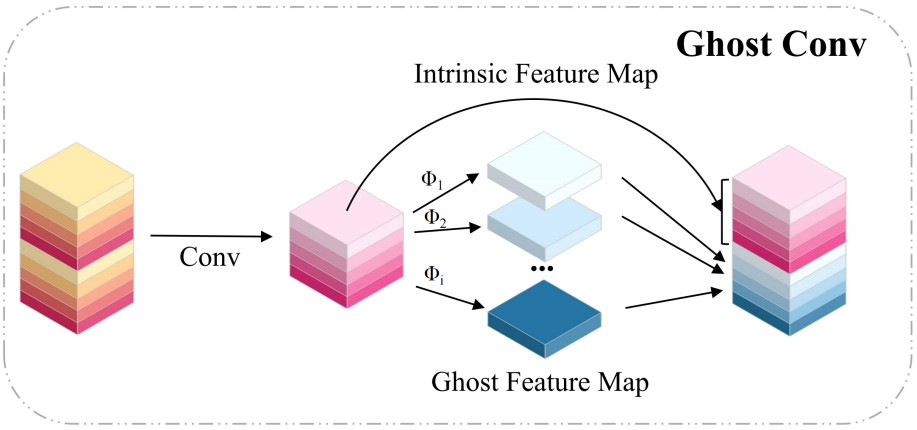


Figure 6. GhostConv Architecture.

2.4.2. Efficient Multi-Scale Feature Extraction Modules: C2f_ConvX and C2f_MConv

Standard convolutions face inherent limitations due to their fixed receptive fields, while existing multi-branch architectures, though capable of providing multi-scale information, often incur substantial computational redundancy. To resolve this "efficiency versus feature diversity" dilemma, we propose an original design philosophy of "asymmetric heterogeneous convolution". Building upon this foundation, we pioneer two novel composite convolution operators (MConv (Figure 7) and ConvX (Figure 8)) and introduce two groundbreaking bottleneck modules (C2f_ConvX and C2f_MConv).

The core design of MConv and ConvX follows the "Split–Process–Concatenate" strategy, a mechanism aimed at efficiently capturing and fusing the inherently multi-scale features in UAV imagery. The key lies in "differentiated processing": unlike standard group convolution, which primarily aims to homogeneously reduce computational load, this strategy intentionally assigns different convolutional operations—varying in kernel scale, shape, or type—to different parallel branches. This design enables each branch to

specialize in extracting specific types or scales of feature information—small kernels for fine texture details, large kernels for contours and macroscopic structures, and asymmetric convolutions potentially more sensitive to edges in specific orientations. Since the features extracted by each branch are designed to be highly complementary, they collectively cover a broader feature spectrum. Consequently, in the final "concatenate" stage, a simple channel concatenation operation itself constitutes an efficient and information-rich feature fusion method, directly pooling diverse information from different receptive fields and semantic levels. The advantage of this mechanism is that it avoids the extra overhead often required by traditional parallel structures (like standard group convolution) to facilitate information interaction, such as the channel shuffle operation needed in ShuffleNet [26], or the additional 1×1 Pointwise Convolution used for channel information fusion in the depthwise separable convolutions of MobileNet [34]. Ultimately, this strategy improves the model's capacity to represent multi-scale, multi-morphology features, concurrently realizing optimized computational efficiency.

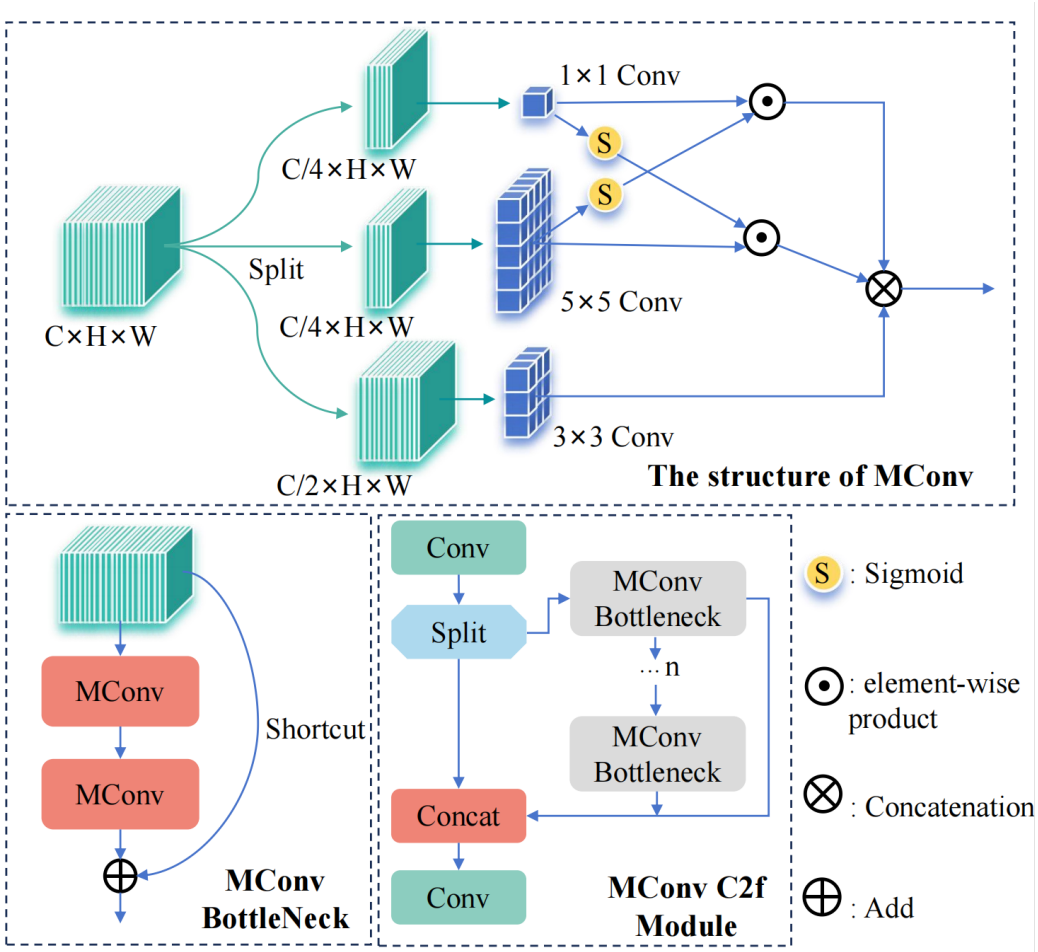


Figure 7. MConv Convolution, MConv Bottleneck, and C2f module architecture based on MConv Bottleneck.

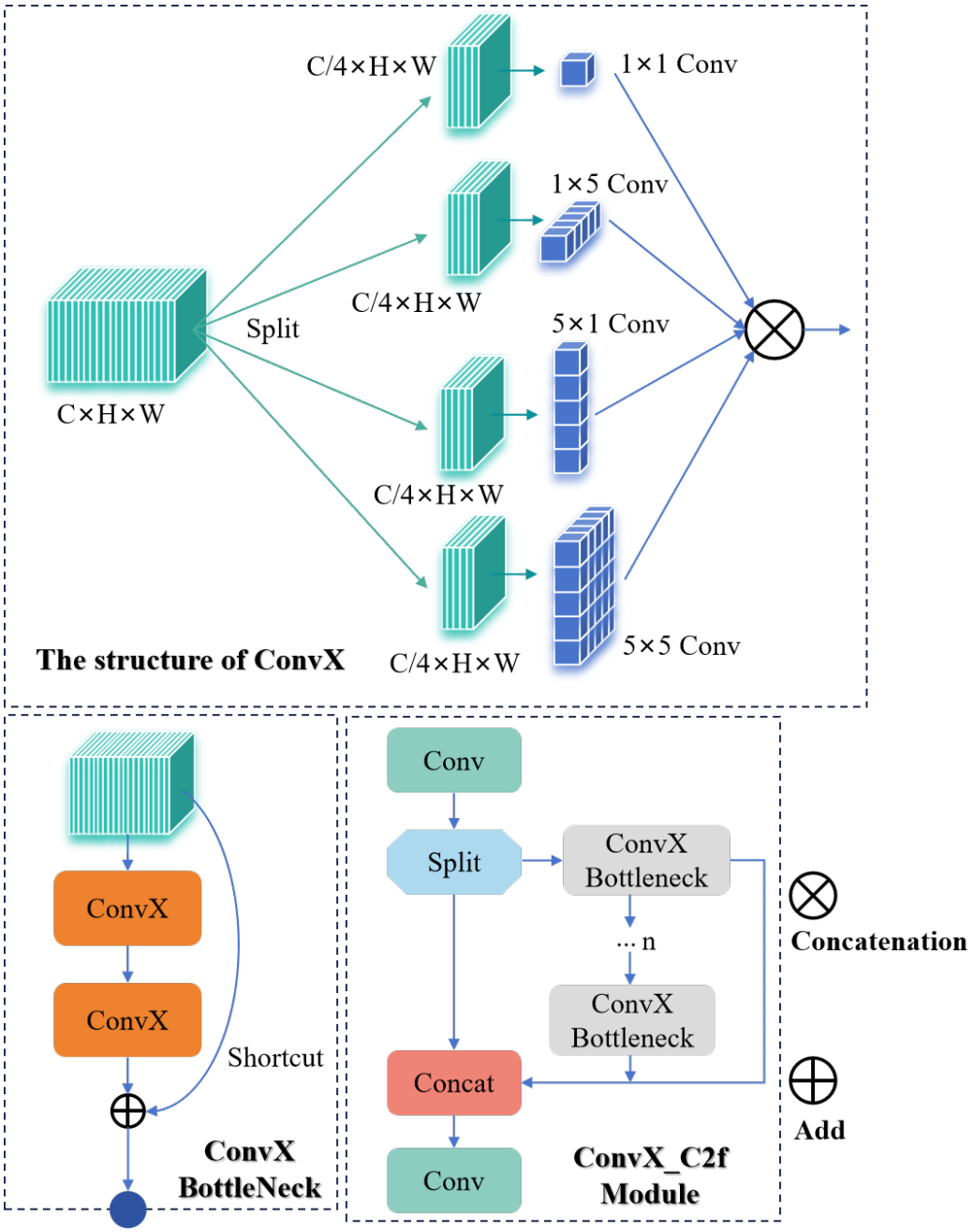


Figure 8. ConvX Convolution, ConvX Bottleneck, and C2f module architecture based on ConvX Bottleneck.

MConv (Figure 7): The input (C channels) is split into three branches of C/4, C/4, and C/2 channels. These are processed by 1 × 1, 5 × 5, and 3 × 3 convolutions, respectively. The output of the 1 × 1 branch, after a Sigmoid activation, is element-wise multiplied with the output of the 5 × 5 branch. Finally, the attention-enhanced features, the 3 × 3 branch's output, and the original output of the 1 × 1 branch undergo concatenation to restore C channels.

ConvX (Figure 8): The input (C channels) is uniformly split into four branches (each C/4). These are processed by 1 × 1 Conv, 5 × 1 Conv, 1 × 5 Conv, and 5 × 5 Conv, respectively. Finally, the four branch outputs undergo concatenation to restore C channels.

Based on MConv and ConvX, we constructed corresponding MConv Bottleneck and ConvX Bottleneck modules and used them to replace the bottleneck blocks in the original C2f module, forming the C2f_MConv and C2f_ConvX modules.

This design offers two prominent advantages: a significant reduction in computational complexity and an effective enhancement of feature representation capability.

1. Significant reduction in computational load: Specifically, taking a feature map of size $C \times H \times W$ as an example, the comparison of computational costs (FLOPs) between standard convolution and MConv/ConvX is as follows:

Standard convolution:

$$\text{FLOPs} = H_{\text{out}} \times W_{\text{out}} \times C_{\text{out}} \times (K^2 \times C_{\text{in}}) \tag{7}$$

where H_{out} represents the height of the output feature map, W_{out} represents the width of the output feature map, C_{out} represents the number of channels in the output feature map, K represents the size of the convolution kernel, and C_{in} represents the number of input channels.

MConv:

$$\text{FLOPs}_{\text{branch1}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 1 \times 1 = \frac{HWC^2}{16} \tag{8}$$

$$\text{FLOPs}_{\text{branch2}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 5 \times 5 = \frac{25HWC^2}{16} \tag{9}$$

$$\text{FLOPs}_{\text{product}} = H \times W \times \frac{C}{4} \tag{10}$$

$$\text{FLOPs}_{\text{branch3}} = H \times W \times \left(\frac{C}{2}\right) \times \left(\frac{C}{2}\right) \times 3 \times 3 = \frac{9HWC^2}{4} \tag{11}$$

$$\text{FLOPs}_{\text{MConv}} = \sum_{i=1}^{3} \text{FLOPs}_{\text{branch}i} + \text{FLOPs}_{\text{m}} = \frac{31HWC^2}{8} + \frac{HWC}{4} \approx 3.875HWC^2 \tag{12}$$

where $\text{FLOPs}_{\text{branch}_i}$ represents the computational cost of the i-th branch, $\text{FLOPs}_{\text{element-wise}}$ represents the computational cost of element-wise multiplication between branch 1 and branch 2, and $\text{FLOPs}_{\text{MConv}}$ represents the computational cost of the MConv composite convolution.

ConvX:

$$\text{FLOPs}_{\text{branch1}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 1 \times 1 = \frac{HWC^2}{16} \tag{13}$$

$$\text{FLOPs}_{\text{branch2}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 1 \times 5 = \frac{5HWC^2}{16} \tag{14}$$

$$\text{FLOPs}_{\text{branch3}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 5 \times 1 = \frac{5HWC^2}{16} \tag{15}$$

$$\text{FLOPs}_{\text{branch4}} = H \times W \times \left(\frac{C}{4}\right) \times \left(\frac{C}{4}\right) \times 5 \times 5 = \frac{25HWC^2}{16} \tag{16}$$

$$\text{FLOPs}_{\text{ConvX}} = \sum_{i=1}^{4} \text{FLOPs}_{\text{branch}i} = \frac{9HWC^2}{4} = 2.25HWC^2 \tag{17}$$

where $\text{FLOPs}_{\text{branch}_i}$ represents the computational cost of the i-th branch, and $\text{FLOPs}_{\text{ConvX}}$ represents the computational cost of the ConvX composite convolution.

As demonstrated in Equations (12) and (17), the computational costs of MConv and ConvX are 3.875HWC^2 and 2.25HWC^2, respectively. Compared to a standard 5×5 convolution (25HWC^2 FLOPs) providing equivalent receptive fields, their computational costs represent merely 15% and 9% of the latter. This substantial theoretical efficiency advantage proves that our heterogeneous design achieves significant computation reduction while

maintaining rich feature representation, thereby effectively meeting real-time requirements for UAV platforms.

2. Enhanced feature richness: Thanks to the diverse and complementary features extracted and fused by the "differentiated processing" strategy, the final output feature map possesses stronger representation capabilities. This helps improve the model's detection accuracy when dealing with targets of varying sizes and blurred details in UAV images.

Finally, to cater to the feature requirements at different positions in the network, we strategically deployed the following two modules:

C2f_MConv module is deployed at the connection point linking the backbone network and the neck network. This position requires effective integration of high-level semantic information from the deep backbone network with shallow-level detail information. The multi-scale parallel processing capability of C2f_MConv can better bridge and fuse semantic and spatial features from different levels, providing a solid foundation for subsequent cross-level feature fusion.

C2f_ConvX is deployed near the detection head. The detection head is directly responsible for predicting target class and location, requiring more detailed and discriminative features. C2f_ConvX, with its diverse convolutional branches, can extract richer and more comprehensive fine-grained features, providing higher-quality information support for the final detection task, thereby improving the accuracy of target localization and classification.

2.4.3. Composite Receptive Field Lightweight Convolution: DRFAConvP Module

While GhostConv provides an effective theoretical framework for model lightweighting, its reliance on static and homogeneous linear transformations to generate "ghost" features inevitably sacrifices feature richness—a critical drawback for detail-sensitive UAV remote sensing tasks. To break this "efficiency–performance" trade-off, we propose DRFAConvP, a novel method that transforms GhostConv's simple linear operations for feature generation into an intelligent feature enhancement process. This innovation integrates two advanced concepts: (1) the multi-branch heterogeneous processing from our composite convolutions, and (2) a dynamic input-dependent spatial attention mechanism inspired by RFAConv [35]. As illustrated in Figure 9, DRFAConvP's core advancement lies in its dynamic attention mechanism and sophisticated heterogeneous transformations, enabling stronger feature representation at low computational cost; the structure of the receptive-field attention mechanism is detailed in Figure 10.

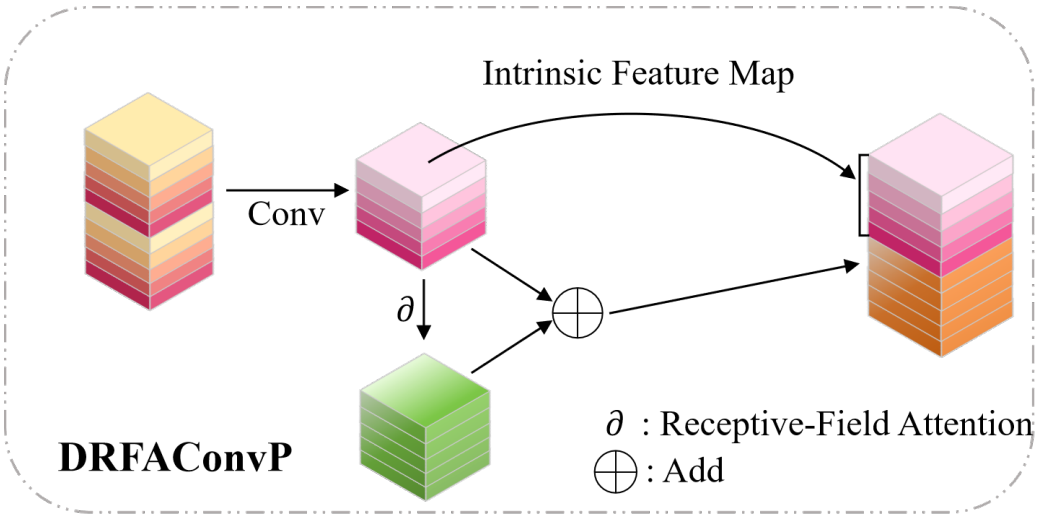


Figure 9. Overall architecture of DRFAConvP.

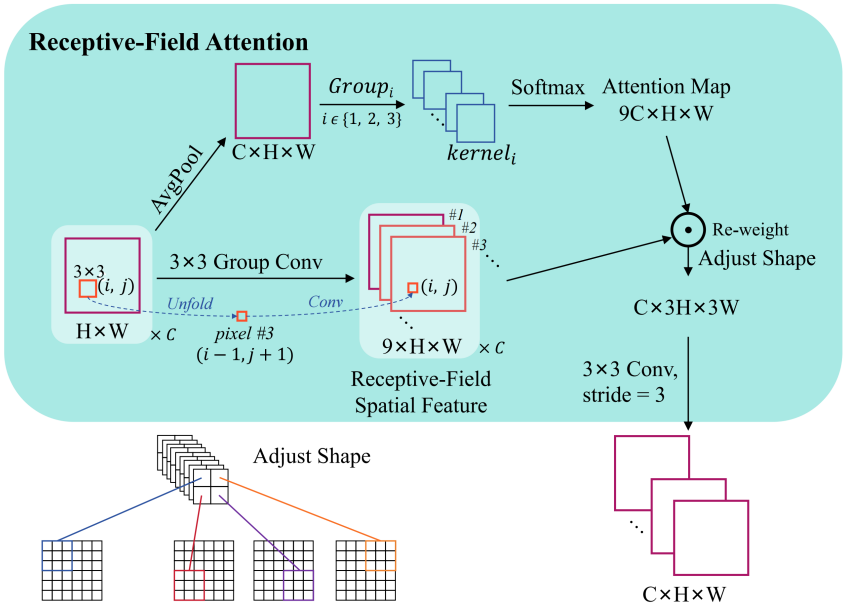


Figure 10. Receptive-field Attention.

Through the above design, the DRFAConvP module, in the process of generating "ghost" features, no longer relies on fixed, simple linear transformations. Instead, it introduces a dynamic receptive field attention mechanism that depends on the input content. Consequently, the convolutional process can adaptively fine-tune the weighting coefficients across distinct spatial zones, guided by the input features. This results in a more efficient acquisition of critical details vital for subsequent tasks, considerably elevating the network's feature descriptive power. Compared to the original GhostConv, DRFAConvP effectively mitigates the problem of accuracy degradation while maintaining model lightweightness and computational efficiency. The detailed procedure of the DRFAConvP module is outlined in Algorithm 1.

Algorithm 1 DRFAConvP Module operations

Input: Input feature map X
Output: Output feature map Z
procedure DRFAConvP (X)
▷ 1. Intrinsic Feature Extraction
$Y \leftarrow \text{Conv1} \times 1(X)$ ▷ Channels of Y are typically half of Z's target channels

▷ 2. Dynamic Weight Generation
$P_H \leftarrow \text{AdaptiveAvgPool}_{\text{Horizontal}}(Y)$
$P_V \leftarrow \text{AdaptiveAvgPool}_{\text{Vertical}}(Y)$
$P_{\text{sum}} \leftarrow P_H + P_V$
$W_{\text{spatial}} \leftarrow \text{GroupConv1} \times 1(P_{\text{sum}}, \text{groups} = \text{channels}(Y))$ ▷ Kernel 1×1
$W \leftarrow \text{Softmax}(W_{\text{spatial}})$ ▷ Spatial dynamic weights

▷ 3. Feature Transformation and Weighting
$F_{\text{transformed}} \leftarrow \text{DWConv}(Y)$ ▷ Depth-wise separable convolution
$F_{\text{weighted}} \leftarrow F_{\text{transformed}} \odot W$ ▷ Element-wise multiplication

▷ 4. Information Fusion and Enhancement
$F_{\text{activated}} \leftarrow \text{ReLU}(F_{\text{weighted}})$
$\text{Conv}F_{\text{weighted}} \leftarrow \text{Conv3x3}(F_{\text{activated}})$
$Y_{\text{enhanced}} \leftarrow Y + \text{Conv}F_{\text{weighted}}$ ▷ Residual connection

▷ 5. Feature Concatenation
$Z \leftarrow \text{Concatenate}([Y, Y_{\text{enhanced}}], \text{axis} = \text{channel})$
return Z
end procedure

2.5. Wise-IoU-Based Loss Function Design

The standard YOLOv8s network uses the CIoU loss to guide its box predictions, measuring differences based on the overlap area, the distance between box centers, and how similar their aspect ratios are. But when we work with UAV datasets—where targets vary greatly in size, backgrounds are cluttered, and infrared images often have low contrast—these less reliable samples can skew training. The CIoU loss may end up focusing too much on large errors from poor-quality labels, or it might give undue weight to anchors that happen to match noisy ground-truth boxes well. Either way, this imbalance can hurt the model's ability to generalize and ultimately worsen its localization performance.

To overcome these challenges and boost localization accuracy on UAV platforms, we employ the Wise-IoU (WIoU) loss function [36] for bounding-box regression. The main innovation of WIoU is its dynamic, non-monotonic focusing mechanism. Rather than using fixed criteria or relying solely on IoU to gauge anchor quality, WIoU introduces an outlier degree β, defined as the ratio between an anchor's instantaneous IoU loss (detached from the gradient graph) and the moving average of IoU losses:

$$\beta = \frac{L_{\text{IoU}}^{*}}{\overline{L_{\text{IoU}}}} \tag{18}$$

where L_{IoU}^{*} denotes the current loss value with gradients blocked, and $\overline{L_{\text{IoU}}}$ is its dynamic mean.

Building on the outlier degree β, WIoU introduces an adaptive gradient modulation strategy. It reduces the gradient contribution from high-quality anchors (i.e., those with smaller β), which are already well-aligned with the ground truth, helping to avoid overfitting. At the same time, it also lowers the gradient for low-quality anchors with larger β, often caused by noisy labels or inherently difficult examples. This dual suppression prevents extreme cases—either too easy or too poor—from dominating gradient updates, which could otherwise impair the model's generalization. Consequently, WIoU encourages the model to prioritize moderate-quality samples, where improvements are most meaningful.

Crucially, this strategy is dynamic: the benchmark value $\overline{L_{\text{IoU}}}$ is updated continuously during training, allowing gradient weights to adapt based on the model's evolving state.

The formulation of the WIoU v3 loss is

$$L_{\text{WIoUv3}} = r \cdot L_{\text{WIoUv1}} \tag{19}$$

Here, L_{WIoUv1} represents the first version of the WIoU loss, enhanced with a distance-based attention mechanism that adjusts the standard IoU loss as follows:

$$L_{\text{WIoU}_{\text{v1}}} = R_{\text{WIoU}} \cdot L_{\text{IoU}} \tag{20}$$

We compute the attention term R_{WIoU} as

$$R_{\text{WIoU}} = \exp\left(\frac{(x - x_{\text{gt}})^2 + (y - y_{\text{gt}})^2}{(W_g^2 + H_g^2)^{*}}\right) \tag{21}$$

This term penalizes predictions that are spatially far from the target center, using the size of the ground truth box to normalize the displacement.

In this context, $L_{\text{IoU}} = 1 - \text{IoU}$ represents the base IoU loss, while R_{WIoU} acts as a spatial attention coefficient. The terms (x, y) and $(x_{\text{gt}}, y_{\text{gt}})$ denote the center coordinates of the predicted and ground truth bounding boxes, respectively. The variables W_g and H_g refer to the width and height of the smallest enclosing box that covers both predicted and

ground truth boxes. The asterisk ($*$) marks that W_g^2 and H_g^2 are detached from gradient computation—treating them as constants to help keep the optimization stable.

The key focusing term r is derived based on the outlier degree β, as defined by

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{22}$$

Here, α and δ are tunable parameters that shape the focusing curve.

In conclusion, WIoU adopts a novel dynamic focusing strategy that adaptively reweights gradient contributions based on anchor box quality, represented by β. This mechanism suppresses the influence of both overly easy and noisy hard examples, redirecting training emphasis toward moderately challenging samples—those with higher potential for meaningful learning. Such a property is crucial when handling UAV imagery, which often contains cluttered scenes and imprecise annotations. By emphasizing more learnable samples, WIoU contributes to improvements in localization precision and generalization capability.

To evaluate the suitability of Wise-IoU (WIoU) for object detection tasks involving infrared UAV images, we compared it with several widely adopted IoU-based loss functions. These include DIoU [37], EIoU [38], SIoU [39], Shape-IoU [40], MPDIoU [41], and Focal-GIoU [42]. All comparisons were conducted on a unified model architecture incorporating the backbone and neck improvements proposed in this paper. The detailed results and performance analysis of this comparison are presented in the Model Training and Evaluation section.

3. Model Training and Evaluation

3.1. Datasets

This study primarily utilizes the HIT-UAV dataset [43] for model design and optimization. Published by Nature Research in April 2023, this dataset is the first high-altitude UAV infrared dataset. It contains infrared images captured by UAVs in various complex environments, such as educational institutions, parking lots, streets, and recreational facilities. Example images are shown in Figure 11. The original dataset annotates five target categories: person, car, bicycle, other vehicle, and 'don't care' regions.

The diverse shooting conditions (e.g., flight altitudes from 30 to 60 m and camera angles from 30 to 90 degrees) result in significant variations in target size and shape, greatly enhancing the dataset's diversity and complexity. This characteristic is crucial for training models, helping to improve their generalization ability and overall robustness to targets of different scales.

In this work, we use a refined subset of the HIT-UAV dataset, which comprises 2898 images with a resolution of 640 $\times$ 512 pixels. The target categories in this subset have been streamlined to three classes: person, bicycle, and vehicle, and it is partitioned into training, testing, and validation sets at a ratio of 7:2:1, respectively. According to statistics, this subset collectively contains 24,751 target bounding boxes. Among these, the vast majority (17,118) are extremely small targets with a size of less than 32 $\times$ 32 pixels (accounting for merely 0.01% of the total image pixels), which poses a significant challenge for small object detection algorithms. The target distribution of the dataset is illustrated in Figure 12.



Figure 11. Example images from the HIT-UAV dataset.

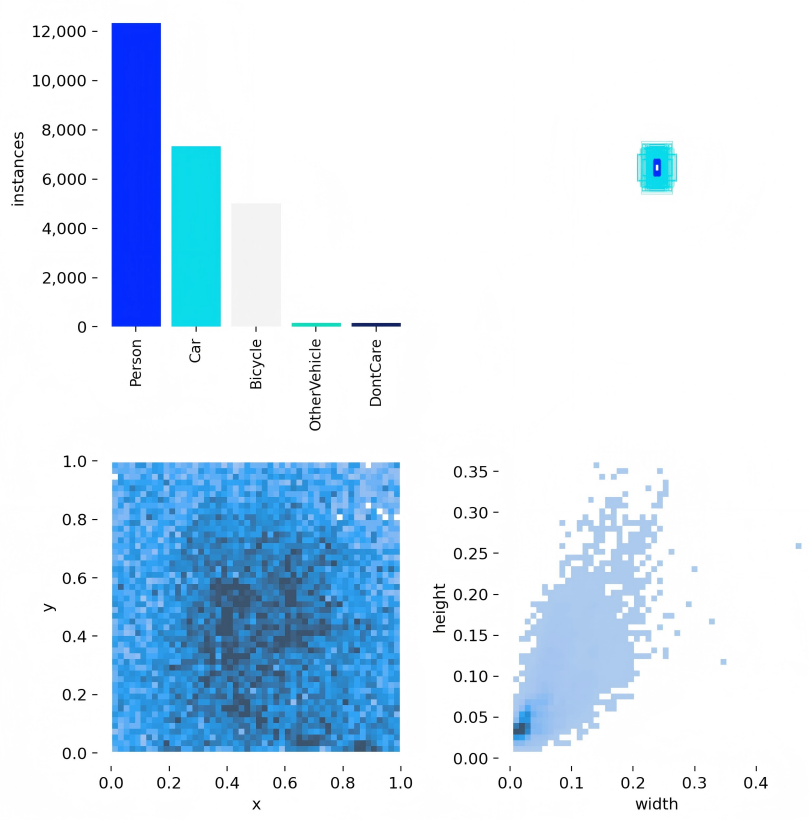


Figure 12. Target distribution of the HIT-UAV dataset.

Overall, the HIT-UAV dataset, with its unique high-altitude infrared perspective and abundant small target samples, provides a valuable resource and a solid foundation for advancing UAV-based infrared object detection and recognition technologies.

3.2. Evaluation Metrics

This paper employs a range of evaluation metrics to assess model performance, including Precision (P), Recall (R), F1 Score, Average Precision (AP), mean Average Precision (mAP), parameter count, and GFLOPs. Among these, the number of parameters and GFLOPs are used to characterize the model's size and computational complexity, respectively. The definitions and formulas for the key metrics are given below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{23}$$

Here, TP (true positive) represents correctly identified positive samples, while FP (false positive) denotes incorrect positive predictions. Precision quantifies the proportion of true positives among all samples predicted as positive.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

FN (false negative) refers to positive instances that were mistakenly predicted as negative. Recall measures the proportion of correctly identified positives out of all actual positive samples.

$$\text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \tag{25}$$

The F1 score is the harmonic mean of Precision and Recall, offering a balanced measure when both false positives and false negatives are of concern.

$$\text{AP} = \int_0^1 P(R)\, dR \tag{26}$$

Average Precision (AP) summarizes the precision-recall trade-off for a specific category by calculating the area under the Precision–Recall $P(R)$ curve.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{27}$$

Mean Average Precision (mAP) is obtained by averaging AP values across N object classes, providing an overall indication of the model's multi-class detection performance.

3.3. Experimental Platform

To assess the performance of the proposed YOLO-SRMX model, we conducted both comparative experiments and ablation studies. The configuration details of the hardware platform used for these evaluations are listed in Table 1.

Table 1. Experimental platform information.

Parameter	Specification
CPU	16 vCPU Intel(R) Xeon(R) Platinum 8481C
GPU	RTX 4090D (24 GB)
RAM	80 GB
Language	Python 3.12 (ubuntu22.04)
Framework	PyTorch 2.3.0
CUDA Version	CUDA 12.1

3.4. Experimental Results

3.4.1. Comparative Analysis of Different Attention Mechanisms

As shown in Table 2, the experiments demonstrate the impact of different attention mechanisms on both detection performance and computational efficiency. Using the baseline model (denoted by $*$) as a reference, its evaluation metrics include a Precision of 0.84, Recall of 0.66, mAP_{50} of 0.760, and mAP_{50-95} of 0.490. The computational complexity is 10.0 GFLOPs, with a parameter count of 11.3 M.

Table 2. Comparative Experiments with Different Attention Mechanisms.

Model	P	R	mAP_{50}	mAP_{50-95}	GFLOPS	Parameters/10^6
$*$	0.84	0.66	0.760	0.490	10.0	11.3
$*$ +GAM	0.76	0.74	0.749	0.463	10.9	12.6
$*$ +iRMB	0.70	0.75	0.749	0.482	11.5	11.4
$*$ +CA	0.85	0.68	0.762	0.483	9.9	11.2
$*$ +ECA	0.82	0.68	0.772	0.482	**9.9**	**11.2**
$*$ +CBAM	**0.87**	0.70	0.785	0.488	9.9	11.3
$*$ +MSDA	0.82	**0.78**	**0.828**	**0.500**	10.0	11.4

Note: '$*$' represents YOLOv8s applying all improvements in this paper except the MSDA attention mechanism. Bold values indicate the best results in each respective column.

After integrating the MSDA attention mechanism, the model achieved notable performance gains while maintaining the same computational cost (10.0 GFLOPs) and incurring only a marginal increase in parameters (11.4 M). Specifically, Recall improved substantially to 0.78, and mAP_{50} rose to 0.828—an increase of 6.8 percentage points—while mAP_{50-95} reached 0.500. Although Precision slightly declined to 0.82, the marked improvements in Recall and both mAP metrics underscore MSDA's effectiveness in enhancing detection capabilities.

In contrast, other attention modules such as GAM and iRMB yielded moderate Recall improvements (0.74 and 0.75, respectively) but recorded lower mAP_{50} values (both at 0.749) and mAP_{50-95} values of 0.463 and 0.482, respectively—underperforming compared to both the baseline and MSDA-enhanced models. Furthermore, GAM increased computational load and model size to 10.9 GFLOPs and 12.6 M, while iRMB also led to higher resource demands (11.5 GFLOPs, 11.4 M).

Lightweight attention modules like ECA and CA reduced the computational burden, but their impact on detection accuracy was limited, with mAP scores falling below those of the baseline. CBAM achieved strong Precision (0.87) and a competitive mAP_{50} of 0.785, with resource usage comparable to the baseline; however, its Recall and mAP_{50-95} remained inferior to MSDA.

Overall, the MSDA attention mechanism offered the best trade-off between performance and efficiency. It delivered significant improvements—particularly in Recall and mAP—without introducing meaningful increases in computational or parameter overhead. As a result, MSDA was adopted as the attention module for the final model configuration.

3.4.2. Comparative Analysis of Different Loss Functions

As summarized in Table 3, the selection of the bounding-box regression loss has a marked effect on model outcomes. For example, the baseline model (denoted by $*$) using the default loss achieves a Precision of 0.762, Recall of 0.751, mAP_{50} of 0.775 and mAP_{50-95} of 0.483. When DIoU is employed, all metrics improve—in particular, mAP_{50} rises to 0.785 and mAP_{50-95} to 0.490—highlighting the benefit of incorporating center-distance considerations. EIoU, SIoU, Shape-IoU, and MPDIoU all significantly improved Precision (P), with EIoU reaching as high as 0.925 and MPDIoU reaching 0.900. Their mAP_{50} values also surpassed the $*$ model, reaching 0.788, 0.782, 0.765, and 0.798, respectively. However,

while these loss functions improved precision, they generally led to a decrease in Recall (R) and mAP_{50-95} metrics. For instance, EIoU's Recall was 0.706, and Shape-IoU's Recall was only 0.689, indicating that they might sacrifice some ability to find targets when pursuing precise localization. Focal-GIoU's Precision and Recall were slightly higher than those of the $*$ model, but its mAP_{50} was 0.770, which was lower than the $*$ model. In contrast, the Wise-IoU loss function adopted in this paper demonstrated the most outstanding performance. It not only boosted Recall to the highest value of 0.780 but also achieved a Precision of 0.820, significantly outperforming the $*$ model and most of the compared loss functions. Notably, the Wise-IoU loss delivers the highest overall effectiveness, with mAP_{50} and mAP_{50-95} reaching 0.828 and 0.500, respectively—substantially exceeding the baseline values of 0.775 and 0.483. These results confirm that WIoU's dynamic, non-monotonic focusing scheme better accommodates fluctuations in sample quality and allocates gradients more judiciously, thereby enhancing localization precision and overall detection performance without sacrificing recall. This validates the choice of WIoU as the optimal loss function.

Table 3. Comparative Experiments with Different Loss Functions.

Model	P	R	mAP_{50}	mAP_{50-95}
$*$	0.762	0.751	0.775	0.483
DIoU	0.781	0.767	0.785	0.490
EIoU	**0.925**	0.706	0.788	0.470
SIoU	0.823	0.723	0.782	0.466
Shape-IoU	0.853	0.689	0.765	0.460
MPDIoU	0.900	0.729	0.798	0.474
Focal-GIoU	0.785	0.737	0.770	0.487
Wise-IoU	0.820	**0.780**	**0.828**	**0.500**

Note: '$*$' represents YOLOv8s applying all improvements in this paper except Wise-IoU. Bold values indicate the best results in each respective column.

To assess the impact of each proposed enhancement in YOLO-SRMX, we adopted YOLOv8s as the reference model and then successively integrated the following modules, including a lightweight backbone network (ShuffleNetV2); an additional small object detection head (More Head); the MSDA attention mechanism; a novel convolution block (New Conv, representing the combined application of DRFAConvP and GhostConv); two efficient multi-scale feature extraction modules (C2f_ConvX and C2f_MConv); and the Wise-IoU loss function. The HIT-UAV dataset was employed for conducting experiments, and the outcomes are summarized in Tables 4 and 5. Each modification resulted in varying degrees of improvement to detection performance and model efficiency, aimed at enhancing model performance for mobile deployment or in resource-constrained scenarios.

Table 4. Ablation Study: Component Configuration.

Experiment ID	ShuffleNetV2	More Head	MSDA	New Conv	C2f_ConvX	C2f_MConv	Wise-IoU
1							
2	✓						
3	✓	✓					
4	✓	✓	✓				
5	✓	✓	✓	✓			
6	✓	✓	✓	✓	✓		
7	✓	✓	✓	✓	✓	✓	
8	✓	✓	✓	✓	✓	✓	✓

Table 5. Ablation Study: Performance Results.

Model	P	R	mAP_{50}	F1 (%)	GFLOPS	Parameters/10^6
1	0.824	0.728	0.768	77	28.8	11.2
2	0.862	0.706	0.776	77	16.1	6.0
3	0.776	0.766	0.774	77	25.4	**5.7**
4	0.866	0.728	0.801	79	25.4	5.8
5	**0.878**	0.707	0.764	77	11.0	12.7
6	0.807	0.736	0.788	77	10.4	11.7
7	0.762	0.751	0.775	76	10.0	11.4
8	0.820	**0.780**	**0.828**	**80**	**10.0**	11.4

Note: Bold values indicate the best results in each respective column.

Table 5 (with setup details in Table 4) clearly demonstrate the progressive contributions of each improvement component to model performance and efficiency. First, replacing the backbone network with ShufflenetV2 (Exp 2) reduced the model parameter count from 11.2 M to 6.0 M, a decrease of 46.4%, and the computational cost (GFLOPS) from 28.8 to 16.1, a reduction of 44.1%. While achieving this significant efficiency improvement, mAP_{50} slightly increased from 0.768 to 0.776, validating the effectiveness of its lightweight structure. Subsequently (Exp 3), adding an extra small object detection head further optimized the parameter count to 5.7 M, although the computational cost increased to 25.4 GFLOPS. This modification significantly improved Recall (R) from 0.706 to 0.766, but mAP_{50} slightly decreased to 0.774.

Next (Exp 4), the MSDA attention mechanism was introduced. With the computational cost remaining at 25.4 GFLOPS and a minor increase in parameters to 5.8 M, mAP_{50} significantly improved from 0.774 to 0.801, an increase of 2.7 percentage points, and the F1 score also rose to 79%. This demonstrates MSDA's ability to enhance feature representation with low resource overhead. Afterward (Exp 5), replacing some traditional convolutions with DRFAConvP and GhostConv (New Conv), drastically reduced the computational cost by 56.7% to 11.0 GFLOPS. However, it also substantially increased the parameter count to 12.7 M and decreased mAP_{50} to 0.764, indicating a trade-off where significant computational efficiency gains negatively impacted model accuracy and parameter size at this stage.

Introducing the C2f_ConvX module (Exp 6) further lowered the computational cost to 10.4 GFLOPS and reduced the parameter count to 11.7 M. Concurrently, mAP_{50} recovered and improved to 0.788, showing that C2f_ConvX effectively enhanced feature fusion capability and optimized model parameter efficiency. Continuing with the introduction of C2f_MConv (Exp 7), the computational cost reached the lowest value in this study at 10.0 GFLOPS, and the parameter count slightly decreased to 11.4 M. However, mAP_{50} slightly dropped to 0.775, suggesting that the main contribution of C2f_MConv lies in optimizing computational complexity. Finally (Exp 8), replacing the loss function with Wise-IoU resulted in significant improvements across key performance metrics: mAP_{50} substantially increased by 5.3 percentage points to 0.828, the F1 score rose to 80%, and Recall (R) reached its optimal value of 0.780. The effective combination of Wise-IoU with the preceding optimization modules contributed to the final model performance optimization.

In summary, compared to the baseline YOLOv8s (Exp 1), the final constructed YOLO-SRMX model (Exp 8) achieved a performance improvement with mAP_{50} increasing by 7.81% (from 0.768 to 0.828) and the F1 score improving by 3.9% (from 77% to 80%). In terms of efficiency, the computational cost (GFLOPS) was substantially reduced by 65.3% (from 28.8 to 10.0), while the parameter count only slightly increased by 1.8% (from 11.2 M to 11.4 M). The ablation experiments individually validated the positive effects of the introduced lightweight structure, attention mechanism, efficient feature fusion modules, and advanced

loss function. The results demonstrate that YOLO-SRMX, through systematic optimization strategies, effectively improves detection accuracy while significantly reducing computational requirements, achieving an excellent balance between accuracy and efficiency, making it highly suitable for application environments with limited computational resources.

3.4.3. Comparative Experimental Results with Different Models

To comprehensively evaluate the performance of the proposed YOLO-SRMX model and clarify its position within existing technologies, we conducted an extensive comparative analysis on the HIT-UAV dataset, comparing it against a diverse range of object detection models. This benchmark set was meticulously curated to ensure a comprehensive evaluation, including the representative two-stage detector Faster R-CNN [8]; the model RT-DRET-X [44], which represents emerging architectural trends; models specifically designed for drone scenarios, such as Drone-YOLO [45] and UAV-DRET [46]; and mainstream YOLO models covering both lightweight and small configurations. This comprehensive selection of models allows for a thorough evaluation of YOLO-SRMX from multiple dimensions, including different design paradigms, application-specific optimizations, and performance–efficiency trade-offs.

The comparative experimental results in Table 6 clearly illustrate the performance and efficiency discrepancies among different models on the drone dataset. The traditional two-stage detector representative, Faster R-CNN, and the new architecture RT-DRET-X both performed poorly on this task, struggling to adapt to the object detection task in infrared images. Their mAP_{50} scores were only 0.562 and 0.405, respectively, far below those of mainstream single-stage models, and their enormous computational costs make them infeasible for deployment on resource-constrained platforms. Secondly, we evaluated models specifically designed for drones: Drone-YOLO and UAV-DRET. Among them, Drone-YOLO demonstrated good computational efficiency with a cost of 12.5 GFLOPS, but its mAP_{50} metric of only 0.772 showed no significant advantage in detection accuracy. UAV-DRET, on the other hand, achieved a mediocre performance in terms of both accuracy and efficiency.

Table 6. Comparison with Different Models on the HIT-UAV Dataset.

Model	Parameters/10^6	Size (MB)	GFLOPS	mAP_{50}	F1 (%)
Faster R-CNN	41.2	137.1	370.2	0.562	53
Drone-YOLO	3.0	6.1	12.5	0.772	77
RT-DRET-X	67.3	129.0	232.4	0.405	38
UAV-DRET	21.5	41.2	73.9	0.749	75
YOLOv5s	9.1	18.1	24.2	0.775	76
YOLOv6s	16.4	32.1	44.9	0.749	75
YOLOv8n	3.2	6.1	8.9	0.786	77
YOLOv8s	11.2	22.0	28.8	0.768	77
YOLOv9s	7.3	15.0	27.6	0.777	78
YOLOv10n	2.8	5.7	8.7	0.743	73
YOLOv10s	8.1	16.2	25.1	0.792	78
YOLO11n	**2.6**	**5.4**	**6.6**	0.768	76
YOLO11s	9.4	18.8	21.7	0.807	78
Ours	11.4	22.7	10.0	**0.828**	**80**

Note: Bold values indicate the best results in each respective column.

The lightweight YOLO model group (including YOLOv8n, YOLOv10n, YOLO11n, etc.) exhibited excellent model efficiency, with parameter counts generally ranging from 2.6 M to 3.2 M and computational loads between 6.6 and 8.9 GFLOPS, facilitating easier deployment. Among these, YOLOv8n showed relatively outstanding performance with an mAP_{50} of 0.786 and an F1 score of 77%, approaching the accuracy levels of some standard-sized

models, but it also possessed the highest model complexity in the lightweight group. The accuracy performance of other lightweight models on this dataset was comparatively average, reflecting that detection accuracy often becomes the main developmental bottleneck when pursuing extreme efficiency.

The small YOLO model group (including YOLOv5s, YOLOv6s, YOLOv8s, YOLOv9s, YOLOv10s, YOLO11s, etc.) typically possesses higher detection accuracy potential than the lightweight models, but this comes at the cost of larger model sizes and higher computational demands, with GFLOPS ranging approximately from 21.7 to 44.9. For instance, YOLO11s achieved the highest mAP_{50} of 0.807 within this group, yet its computational load of 21.7 GFLOPS is still more than double that of the model proposed in this paper.

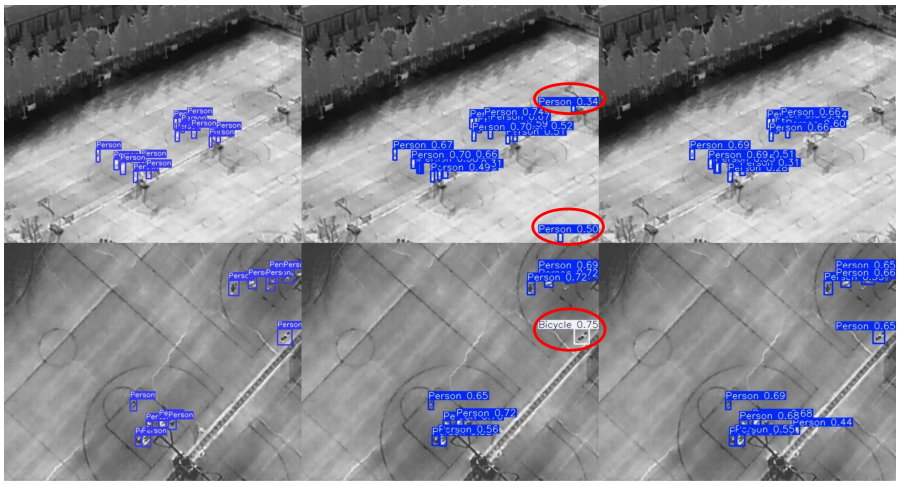
The YOLO-SRMX model proposed in this paper achieves an excellent balance between accuracy and efficiency. Compared to the baseline YOLOv8s, YOLO-SRMX, while having similar parameter counts and model sizes, reduced the computational load (GFLOPS) from 28.8 to 10.0—a substantial decrease of 65.3%. Simultaneously, it significantly improved the mAP_{50} by 7.81% and increased the F1 score from 77% to 80%. Even when compared to the currently advanced small model YOLO11s, YOLO-SRMX achieves an mAP_{50} that is 2.1 percentage points higher and an F1 score 2 percentage points higher, despite requiring less than half the computational load. Although the parameter count of YOLO-SRMX is higher than those of the lightweight models, its computational load is already approaching that group's range while providing markedly superior detection accuracy.

In summary, the comparative experimental results robustly demonstrate the superiority of YOLO-SRMX over current mainstream models, indicating that it represents an efficient and reliable solution for deployment on computationally limited drone platforms tasked with complex aerial detection missions.

3.4.4. Visualization and Comparative Analysis

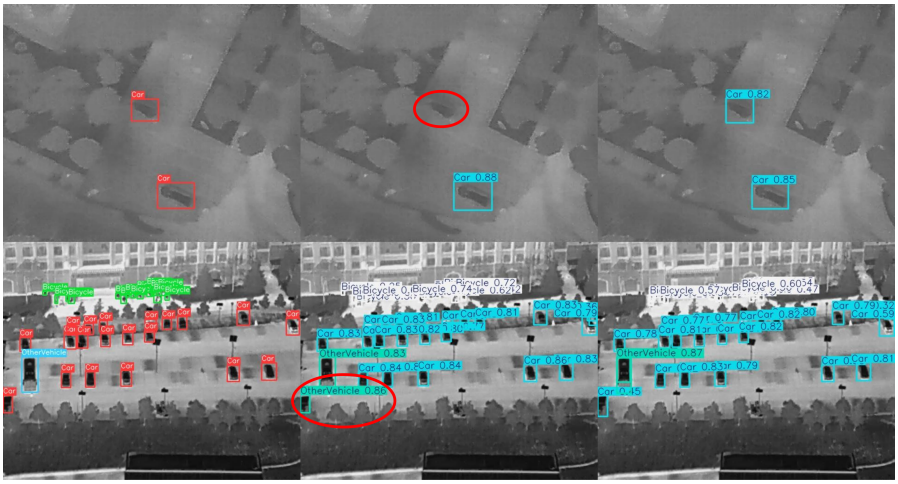
To intuitively validate the detection performance of the proposed YOLO-SRMX model in complex scenarios and to highlight its superiority over the baseline model, YOLOv8s, we have selected typical challenging scenes covering different shooting angles, varying flight altitudes, and diverse target densities for a visual comparative analysis. As shown in Figure 13, each subfigure presents a direct comparison of the detection results.

Through this comparative analysis, it is evident that our model demonstrates significant advantages across several key challenges. First, as depicted in Figure 13a, in scenarios with large-angle oblique shots, the baseline YOLOv8s model suffers from missed and false detections, whereas our model maintains stable detection performance, showcasing superior view invariance. Second, as shown in Figure 13b, when an increase in flight altitude leads to smaller target sizes, YOLOv8s again exhibits missed and false detections, highlighting its limitations in small object detection. In contrast, our model still accurately identifies these small targets. Finally, as illustrated in Figure 13c, in complex situations involving high density and partial occlusion, the performance of the baseline model encounters a significant bottleneck. Specifically, whether facing closely parked vehicles or overlapping targets, YOLOv8s struggles to effectively distinguish them, leading to severe missed and false detections. In contrast, our model, leveraging its superior feature representation capabilities, can penetrate occlusion interference to precisely locate and identify the vast majority of occluded or adjacent targets, demonstrating its excellent robustness in complex and crowded environments.



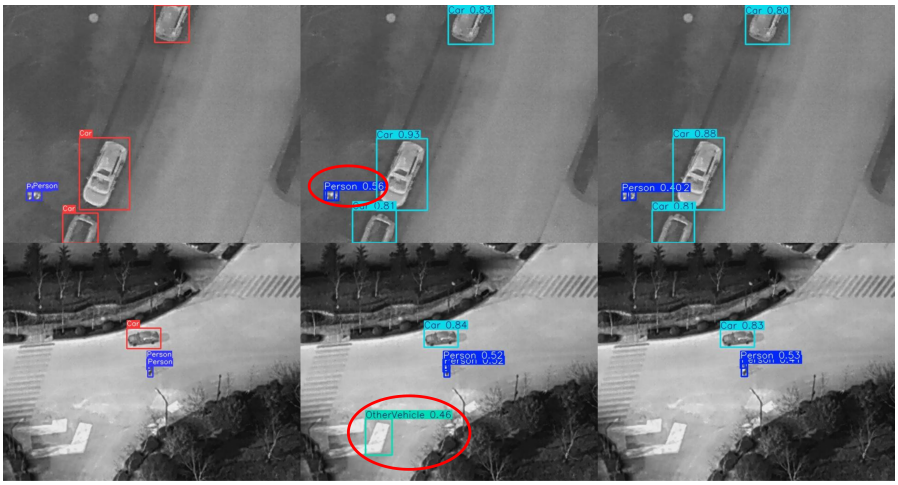
Ground Truth YOLOv8s YOLO-SRMX

(**a**)



Ground Truth YOLOv8s YOLO-SRMX

(**b**)



Ground Truth YOLOv8s YOLO-SRMX

(**c**)

Figure 13. Detection performance of YOLO-SRMX in different environments. Each image displays the results for ground truth (**left**), YOLOv8s (**middle**), and OURS (**right**), with red circles highlighting missed and false detections. (**a**) Detection results under different shooting angles. (**b**) Detection results under different altitude conditions. (**c**) Detection results in high-density scenarios.

3.4.5. Statistical Validation

To comprehensively evaluate the performance stability and effectiveness of the model proposed in this study, we employed a methodology of repeated experiments combined with statistical analysis. We conducted five independent experiments on the model under identical software, hardware, and dataset configurations, meticulously recording its key performance metrics: mean Average Precision (mAP_{50}) and F1 Score.

Statistical analysis of the data revealed that the model achieved a mean mAP@0.5 of 0.821 with a standard deviation of just 0.0075, and a mean F1 Score of 80.8% with a standard deviation of 0.837. This minimal range of fluctuation strongly indicates that the model's performance is not coincidental but rather demonstrates high consistency and reproducibility. Therefore, this series of experiments robustly validates that our proposed model not only achieves excellent detection accuracy but also possesses outstanding performance stability.

3.4.6. Generalization Experiment

To further validate the superiority and robustness of the proposed YOLO-SRMX model in the field of object detection in aerial infrared remote sensing imagery, we conducted a generalization experiment. For this purpose, we fully retrained and evaluated all comparative models involved in this study on the public DroneVehicle infrared image subset.

The data for this experiment were sourced from the infrared image subset of the DroneVehicle dataset [47]. The original DroneVehicle dataset is a large-scale aerial dataset containing an equal number of RGB and infrared images. To maintain consistency with the primary task of this study, we selected only the infrared images from DroneVehicle to construct a subset for this experiment. The infrared subset we constructed includes five vehicle categories: car (26,537 instances), truck (1429), bus (1059), van (570), and freight van (859). This subset was partitioned into a training set of 1799 images, a validation set of 147 images, and a test set of 898 images to assess the generalization capability of each model.

According to the results presented in Table 7, the proposed YOLO-SRMX model outperforms other representative algorithms in both mAP_{50} and F1 Score, including Faster R-CNN, specialized drone detectors (Drone-YOLO, UAV-DRET), and various mainstream YOLO series models. Notably, compared to the baseline model YOLOv8s, YOLO-SRMX's mAP_{50} on the DroneVehicle dataset increased from 0.473 to 0.501, an improvement of 5.9%, while the F1 score also improved by 8.7%. This significant performance enhancement strongly demonstrates that the architecture of YOLO-SRMX is robust and exhibits excellent generalization capability across different data distributions.

Table 7. Comparative experiments of different models on the DroneVehicle dataset.

Model	Parameters/10^6	Size (MB)	GFLOPS	mAP_{50}	F1 (%)
Faster R-CNN	41.2	137.1	370.2	0.430	44
Drone-YOLO	3.0	6.1	12.5	0.431	43
RT-DRET-X	67.3	129.0	232.4	0.446	47
UAV-DRET	21.5	41.2	73.9	0.424	44
YOLOv5s	9.1	18.1	24.2	0.452	43
YOLOv6s	16.4	32.1	44.9	0.471	47
YOLOv8n	3.2	6.1	8.9	0.446	43
YOLOv8s	11.2	22.0	28.8	0.473	46
YOLOv9s	7.3	15.0	27.6	0.462	43
YOLOv10n	2.8	5.7	8.7	0.382	36
YOLOv10s	8.1	16.2	25.1	0.488	48
YOLO11n	**2.6**	**5.4**	**6.6**	0.444	42
YOLO11s	9.4	18.8	21.7	0.486	48
Ours	11.4	22.7	10.0	**0.501**	**50**

Note: Bold values indicate the best results in each respective column.

4. Discussion

The experimental results confirm that the proposed YOLO-SRMX model significantly enhances the accuracy and efficiency of real-time object detection for UAVs, exhibiting outstanding performance, particularly in addressing the challenges of complex backgrounds and small targets commonly encountered from a UAV perspective. The ablation studies clearly demonstrate the contribution of each improved module: the lightweight ShuffleNetV2 backbone, combined with the MSDA attention mechanism, effectively reduces computational costs while enhancing multi-scale feature extraction capabilities, which is crucial for UAV scenarios with complex backgrounds and variable target scales. In the neck network, the combination of DRFAConvP and GhostConv, along with the innovative C2f_ConvX and C2f_MConv modules, successfully balances accuracy and efficiency through efficient multi-scale feature fusion and dynamic receptive field adjustment. Furthermore, the introduction of the Wise-IoU (WIoU) loss function optimizes bounding box regression accuracy and enhances the model's robustness to low-quality samples by intelligently adjusting gradient weights.

A comprehensive analysis reveals that YOLO-SRMX demonstrates robust object detection capabilities from a UAV perspective, especially in complex backgrounds and dense target distributions. Compared to the baseline YOLOv8s and other mainstream models, YOLO-SRMX achieves a superior balance between detection accuracy and computational efficiency (GFLOPs), delivering competitive or even superior detection performance with significantly reduced computational demands. The visualization results further validate the model's robustness under different environmental conditions and intuitively show its superiority over the baseline model in terms of detection accuracy and the reduction of missed detections.

Another important aspect to discuss is the influence of input image resolution on model performance. This study employed a 640×512 resolution for training and optimization. Lowering the resolution would likely degrade accuracy by losing the fine-grained details crucial for identifying small targets, while increasing it would lead to a substantial rise in computational overhead, conflicting with the lightweight design objective of this research. Therefore, the 640×512 resolution represents a reasonable trade-off between detection performance and computational efficiency, aimed at meeting the specific demands of real-time UAV applications.

However, a primary limitation of the current study is that its validation is predominantly based on infrared (grayscale) datasets, which confines the model's verified applications to thermal imaging-related scenarios. A key direction for future work is to extend and validate YOLO-SRMX on color (RGB) remote sensing images to assess its capability in processing richer visual information, thereby adapting it for broader UAV applications, such as standard visible-light monitoring tasks.

In summary, leveraging its significant advantages in accuracy and efficiency, YOLO-SRMX is poised to offer a more powerful and reliable real-time object detection solution for practical UAV applications such as search and rescue, environmental monitoring, and precision agriculture.

5. Conclusions

In this study, we addressed the challenges of accuracy, efficiency, and deployment for object detection in complex UAV environments by proposing a lightweight and highly efficient real-time detection model, YOLO-SRMX. To achieve this, we conducted a deep optimization of the model architecture. First, a novel backbone network was constructed based on the efficient ShuffleNetV2 and integrated with a Multi-Scale Dilated Attention (MSDA) mechanism. This design significantly reduces the model's parameter count while

enhancing its feature-capturing capabilities for targets of varying sizes by dynamically adjusting the receptive field, and it effectively suppresses interference from complex backgrounds. Second, we introduced two major architectural innovations in the neck network. On one hand, we pioneered novel bottleneck modules, C2f_ConvX and C2f_MConv, based on a "split–differentiate–concatenate" strategy. Fundamentally distinct from the homogeneous processing of traditional group convolution, this design decouples richer gradient information and a receptive field hierarchy from the original features at a minimal computational cost by processing asymmetric, multi-type convolutional branches in parallel. On the other hand, our original lightweight convolution, DRFAConvP, radically advances existing concepts by integrating a dynamic attention mechanism into the generation process of "phantom" features, thereby breaking through the performance bottleneck imposed by GhostConv's reliance on static transformations. These original designs work in synergy, greatly enhancing the neck network's adaptability to variable-scale targets from a UAV perspective without sacrificing efficiency. Finally, the Wise-IoU (WIoU) loss function was employed to optimize the bounding box regression process; its unique dynamic non-monotonic focusing mechanism intelligently assigns gradient gains to anchor boxes of varying quality, thereby significantly improving target localization accuracy and the model's generalization ability.

Extensive experimental evaluations of the challenging HIT-UAV and DroneVehicle infrared datasets demonstrate that YOLO-SRMX exhibits significant advantages in both detection accuracy and computational efficiency. Ablation studies clearly showcased the positive contributions of each improved component. On the primary HIT-UAV dataset, compared to the baseline model YOLOv8s, YOLO-SRMX achieved a remarkable performance boost: its mAP_{50} metric increased by 7.81% to 82.8%; the F1 score rose by 3.9% to 80%; and the GFLOPs, a measure of computational complexity, were drastically reduced by 65.3%. Furthermore, comparative experiments against a series of models—encompassing traditional, specialized, and mainstream YOLO architectures—highlighted that YOLO-SRMX achieves a superior balance point between accuracy and efficiency on both datasets. Visualization analyses also confirmed the model's robustness and superior detection performance in complex scenarios involving varied perspectives, flight altitudes, and target densities.

To further enhance the deployment speed and power efficiency of YOLO-SRMX in practical applications, a key future research direction lies in exploring hardware–software co-design. By performing model quantization, pruning, and operator fusion tailored for specific UAV hardware platforms, combined with custom designs that leverage the characteristics of hardware accelerators, it is possible to achieve further acceleration of the model inference process, thereby better satisfying the real-time and high-efficiency object detection demands of UAVs.

Author Contributions: Conceptualization, S.W.; Methodology, S.W.; Software, J.W.; Data curation, H.W.; Writing—original draft, S.W.; Writing—review & editing, H.W., J.W., C.X. and E.Z.; Project administration, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Fundamental Research Funds for the Central Universities under Grant N25GFZ015.

Data Availability Statement: The original HIT-UAV dataset, which was analyzed in this study, is publicly available at https://doi.org/10.1038/s41597-023-02066-6. The specific refined subset of this dataset (including its partitioning into training, validation, and test sets), along with the source code for the YOLO-SRMX model and its trained weights, are available from the corresponding author upon reasonable request.

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

References

1. Ahmad, F.; Qiu, B.; Dong, X.; Ma, J.; Huang, X.; Ahmed, S.; Chandio, F.A. Effect of operational parameters of UAV sprayer on spray deposition pattern in target and off-target zones during outer field weed control application. *Comput. Electron. Agric.* **2020**, *172*, 105350. [CrossRef]
2. Gašparović, M.; Zrinjski, M.; Barković, Đ.; Radočaj, D. An automatic method for weed mapping in oat fields based on UAV imagery. *Comput. Electron. Agric.* **2020**, *173*, 105385. [CrossRef]
3. Nath, N.D.; Cheng, C.S.; Behzadan, A.H. Drone mapping of damage information in GPS-Denied disaster sites. *Adv. Eng. Inform.* **2022**, *51*, 101450. [CrossRef]
4. Kinaneva, D.; Hristov, G.; Raychev, J.; Zahariev, P. Early forest fire detection using drones and artificial intelligence. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019; pp. 1060–1065.
5. Xu, H.; Zheng, W.; Liu, F.; Li, P.; Wang, R. Unmanned aerial vehicle perspective small target recognition algorithm based on improved yolov5. *Remote Sens.* **2023**, *15*, 3583. [CrossRef]
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
12. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
13. Chen, C.; Zheng, Z.; Xu, T.; Guo, S.; Feng, S.; Yao, W.; Lan, Y. Yolo-based uav technology: A review of the research and its applications. *Drones* **2023**, *7*, 190. [CrossRef]
14. Bo, C.; Wei, Y.; Wang, X.; Shi, Z.; Xiao, Y. Vision-based anti-UAV detection based on YOLOv7-GS in complex backgrounds. *Drones* **2024**, *8*, 331. [CrossRef]
15. Wang, W.; Wang, C.; Lei, S.; Xie, M.; Gui, B.; Dong, F. An Improved Object Detection Algorithm for UAV Images Based on Orthogonal Channel Attention Mechanism and Triple Feature Encoder. *IET Image Process.* **2025**, *19*, e70061. [CrossRef]
16. Zhao, S.; Chen, J.; Ma, L. Subtle-YOLOv8: A detection algorithm for tiny and complex targets in UAV aerial imagery. *Signal Image Video Process.* **2024**, *18*, 8949–8964. [CrossRef]
17. Chen, Z.; Ma, Y.; Gong, Z.; Cao, M.; Yang, Y.; Wang, Z.; Wang, T.; Li, J.; Liu, Y. R-AFPN: A residual asymptotic feature pyramid network for UAV aerial photography of small targets. *Sci. Rep.* **2025**, *15*, 16233. [CrossRef]
18. Kakani, V.; Kim, H.; Lee, J.; Ryu, C.; Kumbham, M. Automatic distortion rectification of wide-angle images using outlier refinement for streamlining vision tasks. *Sensors* **2020**, *20*, 894. [CrossRef] [PubMed]
19. Ma, F.; Zhang, R.; Zhu, B.; Yang, X. A lightweight UAV target detection algorithm based on improved YOLOv8s model. *Sci. Rep.* **2025**, *15*, 15352. [CrossRef]
20. Chen, J.; Chen, H.; Xu, F.; Lin, M.; Zhang, D.; Zhang, L. Real-time detection of mature table grapes using ESP-YOLO network on embedded platforms. *Biosyst. Eng.* **2024**, *246*, 122–134. [CrossRef]
21. Pan, P.; Guo, W.; Zheng, X.; Hu, L.; Zhou, G.; Zhang, J. Xoo-YOLO: A detection method for wild rice bacterial blight in the field from the perspective of unmanned aerial vehicles. *Front. Plant Sci.* **2023**, *14*, 1256545. [CrossRef]
22. Li, J.; Chen, M.; Hou, S.; Wang, Y.; Luo, Q.; Wang, C. An improved s2a-net algorithm for ship object detection in optical remote sensing images. *Remote Sens.* **2023**, *15*, 4559. [CrossRef]
23. Liu, L.; Li, P.; Wang, D.; Zhu, S. A wind turbine damage detection algorithm designed based on YOLOv8. *Appl. Soft Comput.* **2024**, *154*, 111364. [CrossRef]
24. Zhang, Y.; Zhang, H.; Huang, Q.; Han, Y.; Zhao, M. DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. *Expert Syst. Appl.* **2024**, *241*, 122669. [CrossRef]
25. Solimani, F.; Cardellicchio, A.; Dimauro, G.; Petrozza, A.; Summerer, S.; Cellini, F.; Renò, V. Optimizing tomato plant phenotyping detection: Boosting YOLOv8 architecture to tackle data complexity. *Comput. Electron. Agric.* **2024**, *218*, 108728. [CrossRef]

26. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Malmö, Sweden, 8–13 September 2018; pp. 116–131.
27. Jiao, J.; Tang, Y.M.; Lin, K.Y.; Gao, Y.; Ma, A.J.; Wang, Y.; Zheng, W.S. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Trans. Multimed.* **2023**, *25*, 8906–8919. [CrossRef]
28. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561.
29. Zhang, J.; Li, X.; Li, J.; Liu, L.; Xue, Z.; Zhang, B.; Jiang, Z.; Huang, T.; Wang, Y.; Wang, C. Rethinking mobile block for efficient attention-based models. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 1389–1400.
30. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
31. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
32. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
33. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
34. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
35. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. RFAConv: Innovating spatial attention and standard convolutional operation. *arXiv* **2023**, arXiv:2304.03198.
36. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.
37. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
38. Zhang, Y.F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [CrossRef]
39. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
40. Zhang, H.; Zhang, S. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv* **2023**, arXiv:2312.17663.
41. Ma, S.; Xu, Y. Mpdiou: A loss for efficient and accurate bounding box regression. *arXiv* **2023**, arXiv:2307.07662.
42. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
43. 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. [CrossRef]
44. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
45. Zhang, Z. Drone-YOLO: An efficient neural network method for target detection in drone images. *Drones* **2023**, *7*, 526. [CrossRef]
46. Zhang, H.; Liu, K.; Gan, Z.; Zhu, G.N. UAV-DETR: Efficient End-to-End Object Detection for Unmanned Aerial Vehicle Imagery. *arXiv* **2025**, arXiv:2501.01855.
47. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]

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.