electronics

Article

Lightweight UAV Object-Detection Method Based on Efficient Multidimensional Global Feature Adaptive Fusion and Knowledge Distillation

Jian Sun [1], Hongwei Gao [2,3,*], Zhiwen Yan [4], Xiangjing Qi [2], Jiahui Yu [5] and Zhaojie Ju [6,*]

[1] School of Graduate, Shenyang Ligong University, Shenyang 110159, China; jiansun6@stu.sylu.edu.cn
[2] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China; 2306610398qxj@stu.sylu.edu.cn
[3] China State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110017, China
[4] Xi'an Modern Control Technology Research Institute, Xi'an 710100, China; bricklin@yeah.net
[5] Department of Biomedical Engineering, Zhejiang University, Hangzhou 310013, China; jiahui.yu@zju.edu.cn
[6] School of Computing, University of Portsmouth, Portsmouth PO1 2UP, UK
* Correspondence: ghw1978@sylu.edu.cn (H.G.); zhaojie.ju@port.ac.uk (Z.J.)

Abstract: Unmanned aerial vehicles (UAVs) equipped with remote-sensing object-detection devices are increasingly employed across diverse domains. However, the detection of small, densely-packed objects against complex backgrounds and at various scales presents a formidable challenge to conventional detection algorithms, exacerbated by the computational constraints of UAV-embedded systems that necessitate a delicate balance between detection speed and accuracy. To address these issues, this paper proposes the Efficient Multidimensional Global Feature Adaptive Fusion Network (MGFAFNET), an innovative detection method for UAV platforms. The novelties of our approach are threefold: Firstly, we introduce the Dual-Branch Multidimensional Aggregation Backbone Network (DBMA), an efficient architectural innovation that captures multidimensional global spatial interactions, significantly enhancing feature distinguishability for complex and occluded targets. Simultaneously, it reduces the computational burden typically associated with processing high-resolution imagery. Secondly, we construct the Dynamic Spatial Perception Feature Fusion Network (DSPF), which is tailored specifically to accommodate the notable scale variances encountered during UAV operation. By implementing a multi-layer dynamic spatial fusion coupled with feature-refinement modules, the network adeptly minimizes informational redundancy, leading to more efficient feature representation. Finally, our novel Localized Compensation Dual-Mask Distillation (LCDD) strategy is devised to adeptly translate the rich local and global features from the higher-capacity teacher network to the more resource-constrained student network, capturing both low-level spatial details and high-level semantic cues with unprecedented efficacy. The practicability and superior performance of our MGFAFNET are corroborated by a dedicated UAV detection platform, showcasing remarkable improvements over state-of-the-art object-detection methods, as demonstrated by rigorous evaluations conducted using the VisDrone2021 benchmark and a meticulously assembled proprietary dataset.

Keywords: UAVs; object detection; small objects; embedded devices

Citation: Sun, J.; Gao, H.; Yan, Z.; Qi, X.; Yu, J.; Ju, Z. Lightweight UAV Object-Detection Method Based on Efficient Multidimensional Global Feature Adaptive Fusion and Knowledge Distillation. *Electronics* **2024**, *13*, 1558. https://doi.org/10.3390/electronics13081558

Academic Editor: Mahmut Reyhanoglu

Received: 5 March 2024
Revised: 26 March 2024
Accepted: 29 March 2024
Published: 19 April 2024

1. Introduction

With the rapid development of the industrial field, unmanned aerial vehicles (UAVs) play a pivotal role in the domains of intelligent transportation [1], precision agriculture [2], and smart cities [3]. UAVs can carry out various activities efficiently because of their affordability, versatility, and portability. Utilizing UAV image object-detection techniques can significantly enhance operational efficiency and minimize human and resource expenses, especially in situations with challenging terrain.

Object detection [4–6] is a fundamental challenge in the field of computer vision, the core of which is to accurately locate and classify objects in images or videos. With the continuous advancement of hardware and computational power, deep learning-based object-detection algorithms have been gradually applied in UAV images by virtue of their strong generalization ability and high detection accuracy. However, UAV images face unique challenges that differentiate them from natural scenes. The substantial variation in UAV flight altitude leads to a higher proportion of small objects, pronounced scale variations, and intricate spatial backgrounds, thereby diminishing the robustness of existing detectors. Furthermore, the limitations of high-resolution images and computational resources pose difficulties in deploying these algorithms on UAV edge devices. Therefore, there is an urgent need for UAV image object-detection models that can maintain high accuracy while being efficiently deployed with minimal latency.

Deep learning-based object-detection algorithms are primarily categorized into one-stage and two-stage approaches. Two-stage detectors, represented by Faster R-CNN [7], exhibit higher accuracy. However, due to the requirement for deeper neural network architectures and more complex model designs, their inference processes are slower, posing a challenge to their efficient deployment on UAV-embedded devices. In contrast, one-stage algorithms, represented by the YOLO series [8], achieve faster inference speeds but sacrifice accuracy compared to models with larger parameters and FLOPs. Recently, Swin Transformers [9] have emerged as effective alternatives to Convolutional Neural Networks (CNNs) [10,11] in UAV image detection tasks. While convolutional layers excel at extracting powerful local features, they are constrained by their limited receptive field, lacking the capacity to model long-range dependencies. On the other hand, Vision Transformers (ViTs) architectures, which rely on the self-attention mechanism, excel in capturing global information and can thus improve global context awareness for UAV image detection. However, the computational complexity of self-attention scales quadratically with image size, which necessitates a significant amount of memory and computational resources, rendering it impractical for some real-world applications of UAV imagery.

In response to the challenge of balancing accuracy with efficiency in UAV image detection, researchers have explored innovative approaches. QueryDet [12] introduces a novel method called Cascaded Sparse Query, which employs a cascaded mechanism. This method predicts the rough positions of small objects at low resolution and then refines detection using high-resolution features for sparse guidance, thereby improving the detection of small objects in UAV images. Du B et al. [13] propose a fast detection method for UAV images based on sparse convolution and adaptive multi-layer masks. This method introduces sparse convolution and a global context-enhancement module to capture global contextual information. It achieves a balance between accuracy and efficiency through the adaptive multi-layer mask module. The method demonstrates significant advantages in computation and inference time, although it suffers from having a large number of model parameters, making deployment on embedded devices challenging. Faster-X [14] presents a lightweight object-detection method for UAVs that is optimized for edge GPU devices. It employs a lighter neck and an auxiliary head structure to enhance the model's inference speed. However, Faster-X does not account for long-range spatial correlations in images, which leads to suboptimal detection performance in complex backgrounds. To address this issue, Wanjie Lu [15] proposes a hybrid detection model that integrates CNNs and transformers. This model utilizes cross-self-attention to capture long-range dependencies at various levels and feeds them into a feature pyramid network for a multi-scale representation. However, this cross-self-attention mechanism still results in the loss of spatial information. Knowledge Distillation [16,17] (KD) is a widely used model-compression strategy, and its core idea is to migrate the advanced knowledge learned from a complex teacher network to a lightweight student network, which enables the student network to have sufficient feature-encoding capability, like the teacher network does. However, there are two limitations in applying KD methods to UAV image target detection. First, existing methods typically enable the student model to learn

only global context information from the teacher network, which limits the reconstruction of local details of the underlying features. Second, MGD [18] employs random marking and AMD [19] uses spatial-attention marking, and these masking strategies focus only on positional information and ignore semantic information, to which UAV images are very sensitive.

To achieve rapid and precise detection of objects in UAV images, our study introduces the Multidimensional Global Feature Adaptive Fusion (MGFAFNET) network. Firstly, we innovatively designed an efficient Dual-Branch Multidimensional Aggregation Backbone Network (DBMA) comprising the Reparameterized Inverted Residual Structure (RIRS) and a Dual-Branch Multidimensional Self-Attention (DMSA). Notably, the RIRS focuses on modeling local relationships in the shallow layers, while the DMSA captures differences among potential dimensions in the deep layers, effectively merging self-attention and cross-attention. This unique approach enhances the representation of deep global features, directing global attention towards dense and occluded regions. Secondly, our study introduces the Dynamic Spatial Perception Feature Fusion (DSPF) to elevate the expression of multi-scale features by aggregating diverse levels of features in a backbone network. This enhancement significantly improves the accuracy with which the system detects small objects. The DSPF incorporates adaptive spatial-weight fusion to elevate the significance of key layer information and mitigate the interference caused by redundant information. Furthermore, the DSPF integrates a feature-refinement module (FRM) to promote feature interaction and incorporates sparse jump connections to facilitate efficient feature reuse and information transfer. Additionally, the integration of a 3D spatial-attention module (3DSAM) utilizing 3D convolutions enables profound exploration of shallow space and extraction of comprehensive information. In addition, our study presents a novel Localized Compensation Dual-Mask Distillation (LCDD) method tailored specifically to UAV images to further refine detection accuracy. The LCDD leverages local and global fusion features from both teacher and student networks, applying feature masks to spatial and channel dimensions. This approach efficiently and comprehensively encodes crucial features and semantic information from the teacher model into the student network, yielding significant improvements in accuracy. Finally, we validate the efficacy of the MGFAFNET on our self-constructed dataset using a UAV object-detection platform. Furthermore, to affirm the model's generalization capabilities, we conduct validation on the publicly available VisDrone2021 dataset. In summary, the primary contributions of this paper are as follows:

(1) To address the challenges faced by UAVs in practical applications, we propose the MGFAFNET, a network specifically designed for UAV images. This method aims to enable devices with limited computational capabilities to meet the requirements of UAV applications in various complex scenarios.
(2) To tackle the complex backgrounds and object-occlusion issues in UAV scenarios, we construct a DBMA Backbone Network. DBMA encodes more effective global feature representations from complex environments and successfully overcomes the high-latency challenges posed by self-attention token mixers.
(3) We introduce a DSPF network that utilizes multi-layer adaptive fusion to suppress irrelevant features. After the data passes through an FRM to enhance multi-scale detection capabilities, the network ultimately incorporates a 3D small-object-detection layer to improve the detection results for small objects.
(4) We design an LCDD distillation method tailored to UAV images. LCDD utilizes the local and global fusion features from the teacher and student networks, adjusting spatial and channel dimensions through adaptive feature masks. The student network comprehensively learns richer encoded information from the larger teacher network.
(5) We constructed a UAV object-detection platform and validated the effectiveness of the proposed method using a self-constructed dataset. Additionally, we performed further validation of the MGFAFNET on the VisDrone2021 dataset [20].

2. Related Work

2.1. UAV Detection Methods

To address the complexity of UAV image object detection, some research efforts have employed techniques that integrate contextual information to enhance the overall adaptability of CNNs. Additionally, certain methods focus on resolving the issue of drastic scale variations in UAV images by incorporating multi-scale feature fusion [21,22]. Another category of methods adopts a two-stage detection strategy that guides the network to learn focused regions first and then refines the processing of these areas [23,24]. This approach effectively addresses challenges related to the detection of small objects and object clustering in UAV imagery. While CNNs excel at leveraging inductive biases to provide prior information for local feature modeling, they are constrained by their attention to local regions, which limits their performance in handling complex scenes and occlusions in UAV images. ViTs [25] rely on multi-head self-attention mechanisms to effectively extract global contextual information, model long-range dependencies, and handle discontinuities in visual features caused by complex backgrounds and occlusions in UAV images. However, the computational complexity of multi-head self-attention is $O\left(N^2\right)$ with respect to the image size, requiring significant memory and computational resources, which is unacceptable for resource-constrained UAV platforms. To reduce the computational complexity of transformer-based structures, the Swin Transformer [9] introduces a hierarchical transformer structure. The backbone network confines self-attention computations to non-overlapping windows and establishes cross-window connections through window shifting, significantly improving the efficiency of self-attention computations. Furthermore, based on the design philosophy of depth-wise separable convolution, researchers have proposed the Separable Vision Transformer (SepViT) [26], which achieves intra-window and inter-window information interaction through depth-wise separable attention, thus resolving the high-latency issue traditionally associated with ViTs.

2.2. KD for Object Detection

In order to develop accurate and lightweight detectors for UAV imagery, researchers have widely applied KD (Knowledge Distillation) methods in recent years to improve the performance of lightweight models. KD, first proposed by Hinton et al. in 2015 [27], achieves performance gains by transferring "dark knowledge" from a high-performance teacher model to a lightweight student model at no additional cost. Current solutions include two main approaches: Logit Mimicking [28] and Feature Imitation [29]. Logit Mimicking enables the logit outputs of student classifiers to be supervised by the logit outputs of teacher classifiers. However, Logit Mimicking operates only on the categorization head, resulting in performance that is limited by the number of categories and that fails to convey positional information. Feature Imitation, which was first proposed for FitNet [30], allows the student model to learn the multi-scale dark knowledge from the teacher model to enhance the consistency of the feature representations between the teacher-student pairs. Although this method extends the KD strategy, it ignores the problem of assigning weights to different target regions. For this reason, L. Zhang [31] applies the attention mechanism to global features by constructing a soft-weighted mask to enhance the information at certain highly important locations. However, L. Zhang's approach focuses only on spatial masking without encoding informative channel clues. DMKD [16] devises a Dual-Masking Knowledge Distillation framework that captures both spatial and channel-wise informative clues for comprehensive masked feature reconstruction. This approach does not consider the comprehensive masking of local features, while local information is crucial for UAV images. For this reason, we propose a Localized Compensation Dual-Mask Distillation (LCDM) based on DMKD. LCDM combines the local and global features of both the teacher and the student and uses spatial and channel attention mechanisms to guide the respective weighted mask features. By doing so, LCDM aims to reconstruct the mask features at both spatial and channel levels effectively.

3. Methodology

3.1. UAV Object-Detection Platform

We constructed a UAV object-detection platform equipped with vision sensors and utilized the platform to evaluate the comprehensive performance of different detection algorithms in a real scenario. UAV object-detection platforms have the advantages of high mobility, low cost, and easy deployment and can perform reconnaissance work in harsh and complex environments. The platform mainly consists of a quadcopter UAV and ground-based information-receiving equipment. The UAV platform can be manually controlled by a remote control or programmed to automatically fly along a predefined route planned by the data-processing unit. According to user settings, the flight line's route is input first, and then the flight-control module guides the UAV to collect data along the planned route. The UAV transmits the flight data and images back to the ground processing module through the digital and image-transmission modules, and the ground processing module receives the data in real time and displays it on the software. After the reconnaissance is completed, the UAV automatically returns to the ground. Subsequently, we implement our proposed algorithm using the PyTorch framework and train it on the Ubuntu system. The trained weight files are converted to the ONNX format and then to TensorRT to accelerate inference. Finally, we deploy the models to the embedded systems and the ground side to meet different requirements. Our platform can not only monitor the UAV flight status and images in real time but also efficiently accomplish the object-detection task. The UAV object-detection platform is shown in Figure 1.

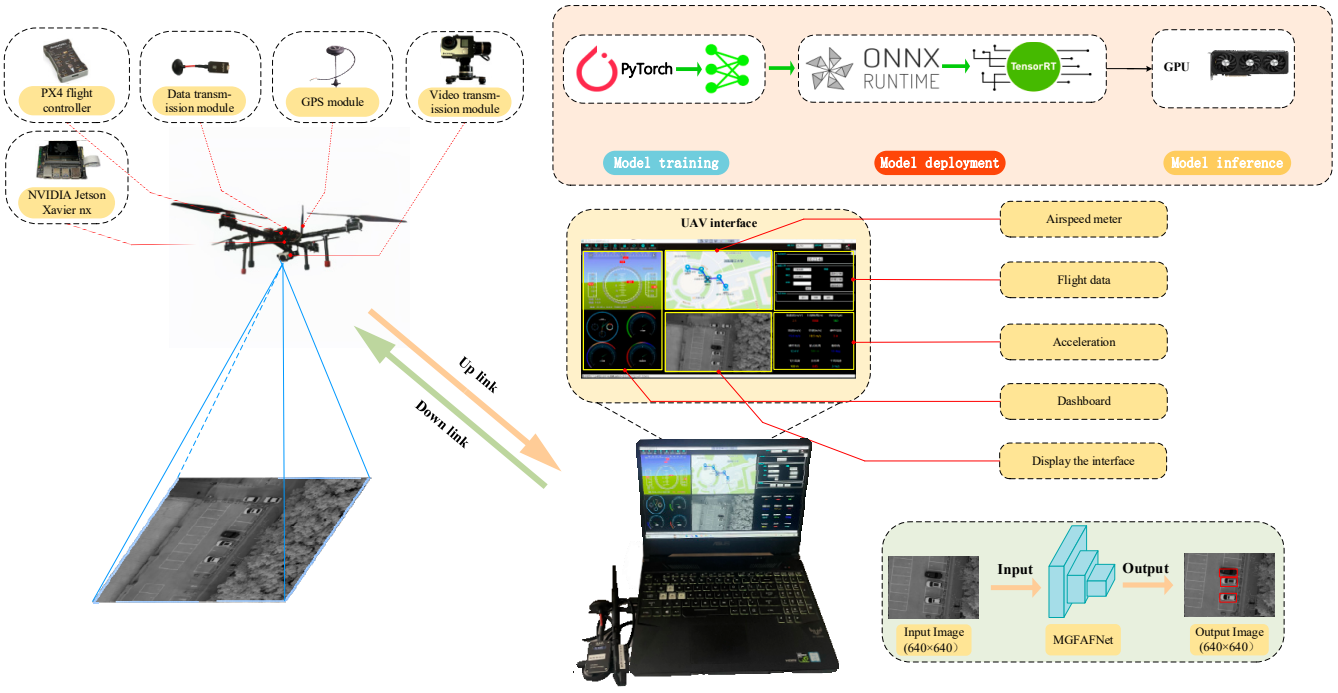

Figure 1. A brief overview of UAV object-detection platforms.

3.2. MGFAFNET

We propose a new lightweight end-to-end object-detection method for UAVs, MGFAFNET, to achieve efficient object detection in complex scenarios. Our network mainly consists of three core components: the DBMA backbone, the DSPF neck, and the LCDD distillation method. DBMA is used to extract global and local shared information from images to improve detection of complex and occluded objects; DSPF is used to fuse backbone network features to improve multi-scale coding capabilities; and LCDD is used to learn features from large model representations to improve network detection accuracy. The

overall structure of our proposed MGFAFNET is shown in Figure 2, and we will introduce the related key components and modules in detail in the following sections.

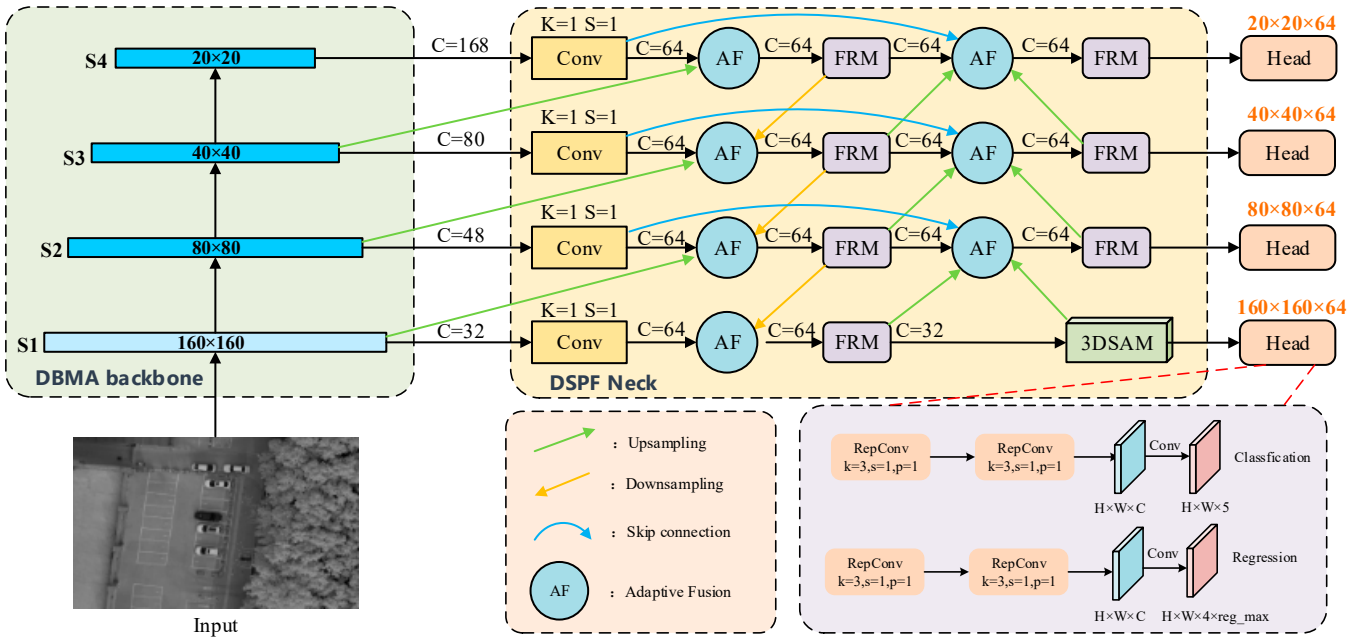

Figure 2. The MGFAFNET architecture.

3.3. DBMA Backbone

The DBMA backbone network inherits the advantages of both CNN and Transformer, achieving high-precision detection of objects against complex backgrounds. DBMA utilizes four stages to generate features at varying scales. In the first two stages, we employ RIRS to extract low-level features, while in the latter two, we integrate RIRS and DMSA to capture both local and long-range dependencies. Our backbone network chiefly comprises the Stem, Deep Down-sampling, RIRS, and DMSA.

Stem: The transformer model typically slices the input image into non-overlapping patches using a large convolution as a stem; however, this practice can lead to the loss of 2D spatial features and edge information. To overcome this limitation, we initially apply a 3×3 convolution with a stride of 2 and an output channel count of 24 for down-sampling. Subsequently, we integrate a 3×3 depth-wise convolution (DW convolution) with a 1×1 convolution to more effectively capture local information. Additionally, we embrace the concept of structural reparameterization to mitigate information loss by utilizing a multi-branch topology during training, which is then discarded during inference to facilitate wider network. This approach significantly reduces the computational load in the inference phase while enhancing network performance.

Deep Downsampling: In ViTs, down-sampling is typically implemented through a separate patch merging layer; however, this approach can lead to the loss of spatial information. To alleviate this problem, we have designed a deep down-sampling module to mitigate the loss of spatial dimensional information resulting from reduced resolution. A schematic of the deep down-sampling module is shown in Figure 3. We first employ a 1×1 convolution to expand the channel dimension, followed by a GELU activation function. We then apply a 3×3 DW convolution for the purpose of down-sampling. The 3×3 DW convolution effectively enhances the network's capability to model local features while maintaining a lightweight structure. Subsequently, the GELU activation function is applied. Finally, we utilize a 1×1 convolution to refine the channel dimension.

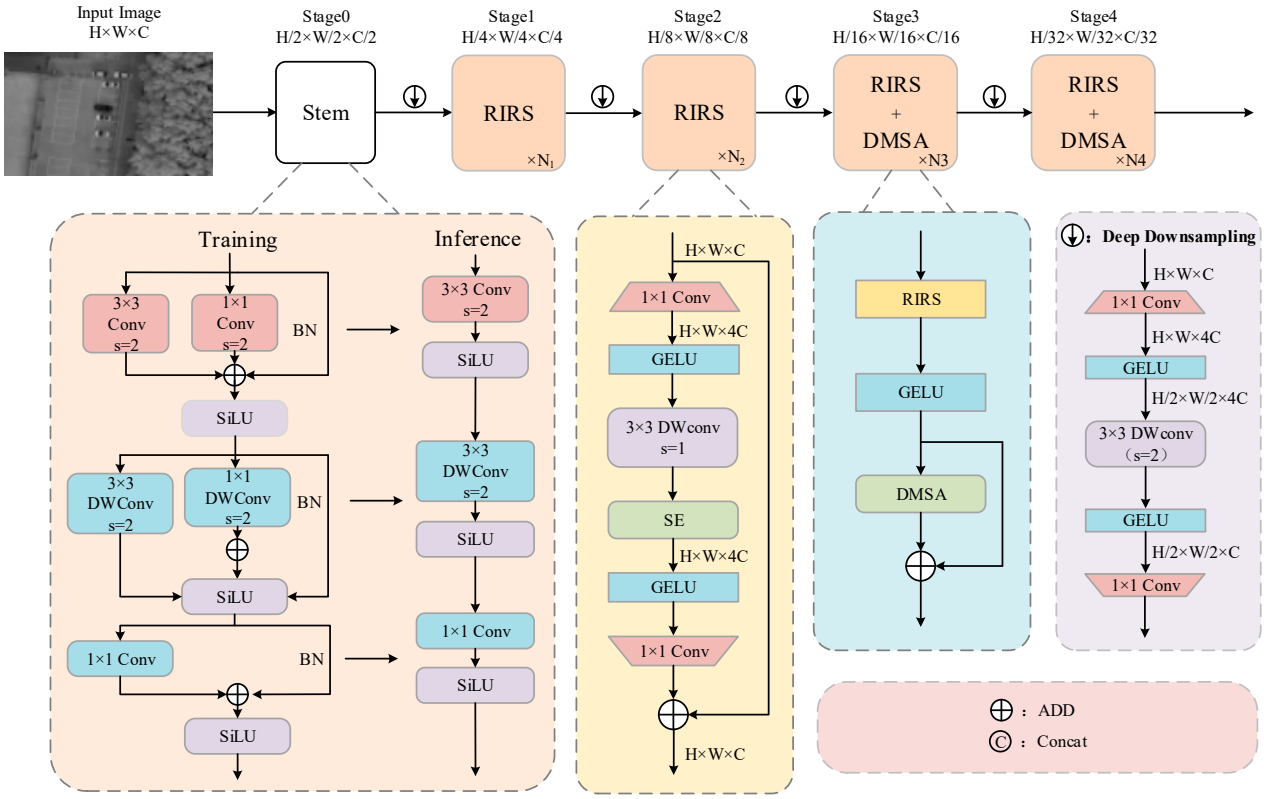

Figure 3. The structure of the DBMA backbone.

RIRS: Inspired by the architecture of MobileNetV3 [32], we propose a RIRS module. RIRS achieves efficient inter-channel interaction through a combination of 1×1 extended convolution and 1×1 projection layers. Notably, the RIRS design incorporates the SE (Squeeze-and-Excitation) layer in only the last two stages of the model to compensate for the limitations of standard convolutions with respect to data-driven attribute processing. This choice is primarily influenced by the greater significance of channel attention mechanisms in processing deeper features.

Given the input token tensor $X_{in} \in R^{C \times H \times W}$, the channel is first expanded by a factor of 4 using a 1×1 expand convolution, followed by an activation function to introduce nonlinearity. Subsequently, spatial features are captured with a 3×3 reparameterized DW convolution, and then the features are projected back to the original channel size using another 1×1 convolution. Afterward, batch normalization is applied to prevent gradient loss. The formula is as follows:

$$Conv_{1\times1}(GELU(SE(DWconv_{3\times3}(GELU(Conv_{1\times1}(X_{in})))))) + X_{in} \quad (1)$$

DMSA: To overcome the inherent shortcomings of conventional attention mechanisms, we have developed a lightweight DMSA, as illustrated in Figure 4. This module combines cross-attention with a lightweight global attention mechanism to effectively capture global contextual information from various dimensions. Following this step, our proposed method is detailed with rigorous mathematical formulations.

For a token input $X \in R^{C \times H \times W}$, where H and W denote the length and width, respectively, and C denotes the number of channels, we compute the query (Q), key (K), and value (V) projections using three linear layers. Self-attention can be represented as follows:

$$MSA(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where K^T denotes the transpose of matrix K and d_k is the dimension of the head. The Softmax is the normalization operation used to compute the attention weights. The original self-attention obtains a global receptive field by establishing associations between all input tokens. However, it has a quadratic computational overhead of O (N^2), resulting in a high computational cost. To mitigate the quadratic complexity of self-attention, we use an efficient cross-attention mechanism to aggregate local dependencies with window priors. The DMSA module is illustrated in Figure 4.

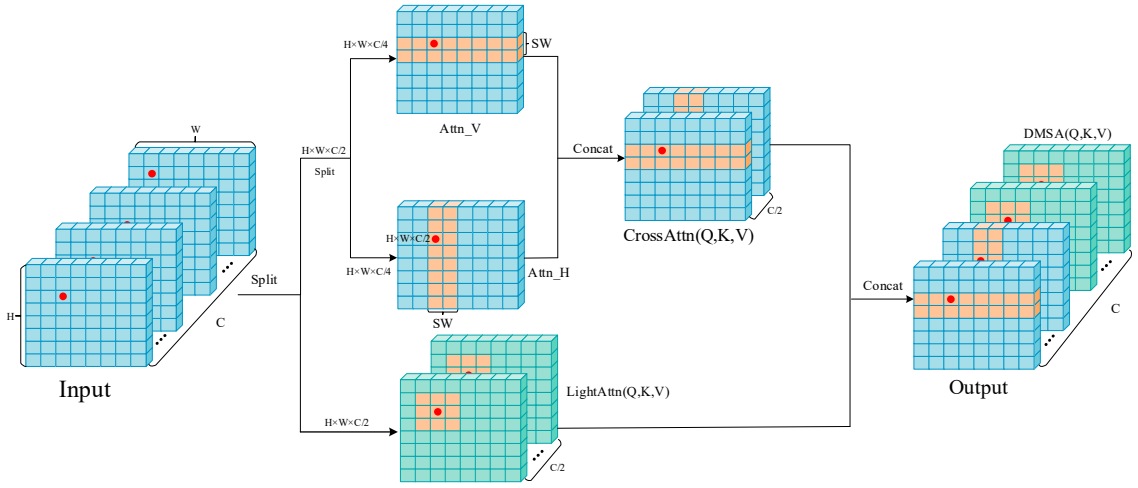

Figure 4. Illustration of the DMSA module.

The cross-attention forms horizontal and vertical stripes by dividing the input data vertically and horizontally and performs the self-attention computation in parallel inside the horizontal and vertical stripes. Specifically, given an input $X \in R^{H \times W \times \frac{C}{2}}$, cross-attention divides the input into two parts $X_h \in R^{H \times W \times \frac{C}{4}}$ and $X_w \in R^{H \times W \times \frac{C}{4}}$ in the channel dimension, and then divides the height and width of the features into non-overlapping stripes of size sw (sw is set to 7) along the vertical and horizontal directions, respectively, with the number of vertical and horizontal stripe attentions $\frac{H}{sw}$ and $\frac{W}{sw}$, respectively. After attention computation has been performed, the stripes in different dimensions are spliced together. $Q, K, V \in R^{N \times d}$. The cross-attention computation formula is shown below.

$$Attn\_H\left(Q_i^H, K_i^H, V_i^H\right) = \left[Softmax\left(\frac{Q_i^H}{\sqrt{d_k}} \cdot \left(K_i^H\right)T\right) \cdot V_i^H + LePe\left(V_i^H\right)\right]_{i=1:H/sw}$$

$$Attn\_V\left(Q_j^v, K_j^v, V_j^v\right) = \left[Softmax\left(\frac{Q_j^v}{\sqrt{d_k}} \cdot \left(K_j^v\right)T\right) \cdot V_j^v + LePe\left(V_j^v\right)\right]_{j=1:W/sw} \quad (3)$$

$$CrossAttn(Q, K, V) = concat(Attn_H, Attn_V)$$

where $Attn\_H$ and $Attn\_V$ denote the output of self-attention on the horizontal and vertical axes, respectively. In the formula, i and j denote the numbers of horizontal and vertical stripes, respectively $LePe$ performs locally enhanced positional encoding on V using a 3 × 3 convolution to learn a more refined local representation. Finally, the outputs of horizontal and vertical self-attention are combined to obtain $CrossAttn$.

Although cross-attention is efficient, it limits the global interactions between different tokens. To efficiently inherit the advantages of global attention, we introduce lightweight global attention to capture important information in UAV images from different dimensions. The lightweight global attention down-samples the $K \in R^{N \times d}$ and $V \in R^{N \times d}$ spatial dimensions using average pooling before performing the attention operation to mitigate the computation overhead and obtain the down-sampled $K', V' \in R^{\frac{N}{4 \times d}}$. In addition, we

also use *LePe* for local position encoding. The lightweight global self-attention can be calculated as

$$LightAttn(Q,K,V) = Softmax\left(\frac{(QK')^T}{\sqrt{d_k}}\right)\cdot V' + LePe\Big((V')^H\Big) \tag{4}$$

Finally, we lightweight global attention and cross-attention to obtain multidimensional attention. The formula is shown below.

$$DMSA(Q,K,V) = concat(CrossAttn(Q,K,V), LightAttn(Q,K,V)) \tag{5}$$

3.4. DSPF Neck

Multi-scale features play a crucial role in object detection. The feature pyramid adopts a top-down model, which facilitates the exchange of information between different scale features and achieves satisfactory results in natural scenes. However, we believe that for the fusion of multi-scale features from UAV images, relying solely on such a paradigm may prove insufficient. The drastic changes in target scale due to the variations in UAV flight altitude and camera angle, as well as the significant increase in the number of small targets, require the detection model to have a more robust multi-scale object detection capability. To this end, we propose a novel approach termed DSPF, distinguishing it from existing multi-scale feature fusion schemes, to enhance multi-scale detection performance in UAV scenarios.

Specifically, DSPF adopts a bidirectional fusion paradigm that utilizes the top-down branches to provide high-level semantic information to bottom-level features, while the bottom-up branches offer spatial location information to deeper features, making the information flow more delicate than that of FPN [33]. To enhance feature multiplexing and to maintain local detail transmission, DSPF adds skip connections to enable more efficient information delivery to the deeper layers. Additionally, DSPF achieves more efficient transfer of semantic and spatial information by aggregating three layers of multi-scale features, as opposed to the layer-by-layer transfer seen in FPN. Moreover, DSPF introduces an adaptive spatial fusion mechanism to improve the anti-interference ability of the key layer information. To ensure that the information after adaptive fusion is enriched, DSPF introduces the Feature Refinement Module (FRM) to further optimize the feature representation. Finally, DSPF employs a 3DSAM, which uses 3D convolution to explore the intrinsic connection of spaces on different scales and, consequently, further improve multi-scale feature detection, especially for small objects.

The DSPF is illustrated in Figure 5. Given three neighboring multi-scale features, $S_{i-1} \in R^{\frac{H}{2}\times\frac{W}{2}\times C}$, $S_i \in R^{H\times W\times C}$, and $S_{i+1} \in R^{2H\times 2W\times C}$, these three-layer features are resampled to the same scale as S_i and concatenated. Then, the dimensionality of the fused features is reduced to 3 by a 1×1 convolution. Softmax operations are applied to the three layers of features to determine the feature weights of each layer: S'_{i-1}, S'_i, and S'_{i+1}. The weights of each layer are multiplied with the corresponding input layer to obtain the adaptive fusion features of each layer: $\hat{S}_{i-1}$, $\hat{S}_i$, and $\hat{S}_{i+1}$. Finally, the three layers of features are summed up to obtain the adaptive fusion features S. The mathematical representation of the entire fusion strategy is as follows:

$$\begin{aligned}\left(S'_{i-1}, S'_i, S'_{i+1}\right) &= Split(Softmax(Conv_{1\times1}(\phi(Up(S_{i-1}), S_i, Down(S_{i+1})))))\\ S &= Conv_{1\times1}\left(S'_{i-1}\otimes Up(S_{i-1}) + S'_i\otimes S_i + S'_{i+1}\otimes Down(S_{i+1})\right)\end{aligned} \tag{6}$$

where $S'_{i-1} \in R^{H\times W\times 1}$, $S'_i \in R^{H\times W\times 1}$, and $S'_{i+1} \in R^{H\times W\times 1}$ are the three layers of features to be fused, respectively, and ϕ is the concatenation operation. The Split denotes separation of channels. $Up()$ and $Down()$ denote two-fold upsampling and downsampling, respectively. S is the feature after adaptive fusion.

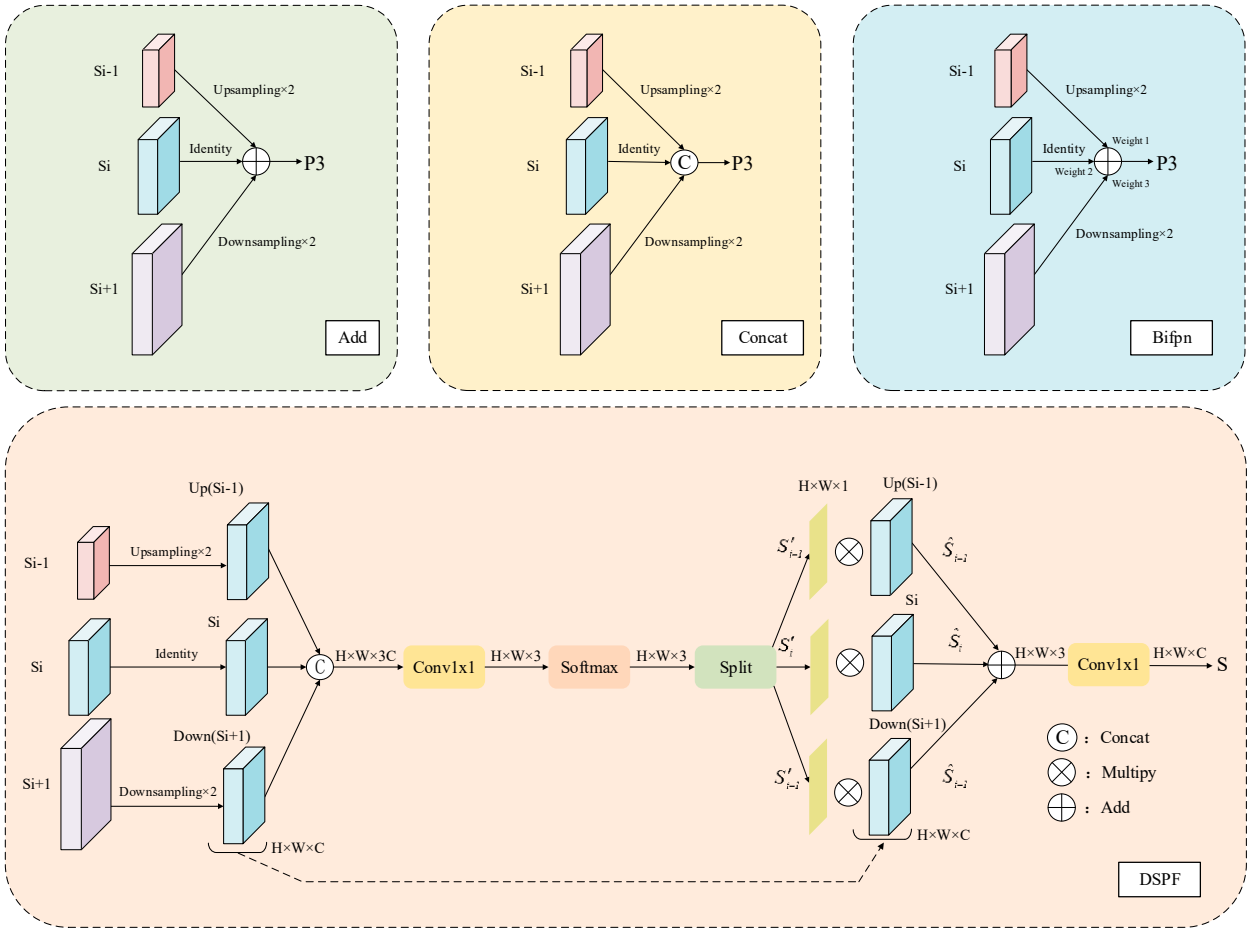

Figure 5. Comparison of different fusion strategies.

FRM: Then, the fused features are then fed into the FRM module, as shown in Figure 6. FRM utilizes reparameterized depthwise convolution to enhance the exchange of information at different scales. The TFRM formulation is shown below:

$$Y = Conv_{1\times1}(\phi(Conv_{1\times1}(x), Conv_{3\times3}(\mathrm{Rep}\,DW_{3\times3}(Conv_{1\times1}(x) + Conv_{1\times1}(x))))) \quad (7)$$

where $X \in R^{H\times W\times C}$ is the input features and $Conv_{1\times1}$ and $Conv_{3\times3}$ denote 1×1 and 3×3 convolutions, respectively. $RepDW_{3\times3}$ refers to 3×3 reparameterized depthwise convolution, and Y denotes the output feature.

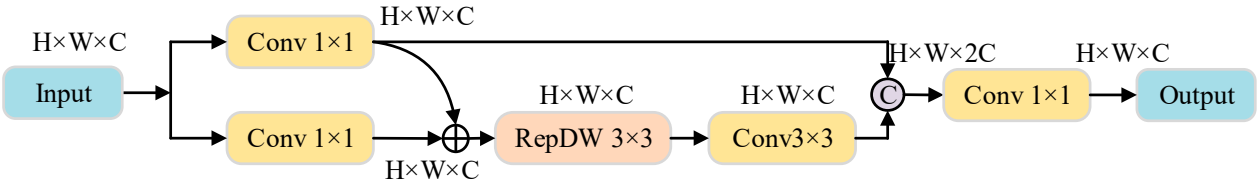

Figure 6. The structure of the FRM.

3DSAM: In addition, we introduce a 3DSAM that utilizes 3D convolution to extract more precise positional information along the scale-space axis, thereby significantly enhancing the perception of small object positions. Figure 7 depicts the 3DSAM. Given a shallow feature input, denoted as Input $\in R^{H\times W\times C}$, we initially employ an 'Unsqueeze' operation to insert an additional dimension along the second axis. Subsequently, a 3D convolution is applied to extract high-dimensional features, which is followed by average pooling to

reduce the features. Afterwards, the 'Squeeze' operation is then utilized to restore the original dimensionality. The input features, after being processed by the Sigmoid function, are element-wise multiplied with the extracted features to obtain the final output results. The 3DSAM structure is shown in Figure 7.

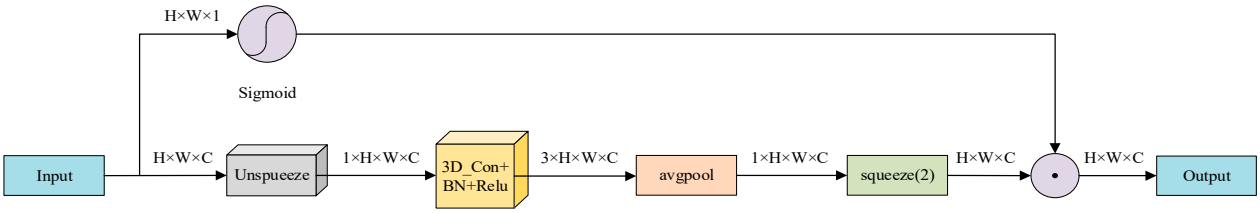

Figure 7. The structure of the 3DSAM.

3.5. LCDD Knowledge Distillation

We propose an LCDD method that utilizes local and global fusion features of the teacher and student networks for spatial and channel feature masking. The LCDD distillation method consists of three key steps: first, acquisition of local and global dual fusion features; second, attention dual thresholding coding using the fusion features; and third, spatial and channel dimensional feature reconstruction. In this process, LCDD first slices the features into patches to compensate for the learning of local information in UAV images. Then, the resulting fused features are used for local and global feature double masking for both the teacher and student networks, enabling the student network to learn more efficiently about the complex feature representations of the teacher network during the processing of UAV images. Finally, feature reconstruction is performed in the spatial and channel dimensions, and losses are constructed using SSIM to ensure that the reconstructed features are similar to the original features in both space and channel. This distillation approach effectively and comprehensively encodes the underlying features and semantic information in the UAV images into the student network, enabling it to fully learn the ability to extract complex and weak information from the UAV images at a level comparable to that of the teacher network.

To obtain the attention features of teacher and student networks in both local and global space, we employ a Softmax function that normalizes the absolute value X_k of each pixel over the channel. At the same time, a temperature parameter τ is introduced to adjust the distribution. LCDD utilizes $P(X)$ to divide the input feature X into n blocks and then performs local attention computation within each block. Subsequently, these local features are concatenated. Finally, the spatial attention feature T_{space} is obtained by summing the local and global attention and averaging the values. The computational formula is detailed below:

$$A_{S/T}(X) = C \cdot Softmax\left(\frac{\frac{1}{C}\sum_{k=1}^{C}|X_k|}{\tau}\right) \tag{8}$$

$$P(X) = \left(X^1 \dots X^n\right) \tag{9}$$

$$L_{space}(P(X_s), P(X_t)) = \varphi\left(A_S\left(X_s^1\right) \dots A_S(X_s^n)\right) + \varphi\left(A_T\left(X_t^1\right) \dots A_T(X_t^n)\right) \tag{10}$$

$$G_{space}(X_t, X_s) = A_T(X_t) + A_S(X_s) \tag{11}$$

$$T_{space} = \frac{L_{space}(P(X_s), P(X_t)) + G_{space}(X_t, X_s)}{2} \tag{12}$$

where $A_{S/T}(\ )$ represents a spatial attention operation applied to the student or teacher features. X_s and X_t represent the input features of the student and teacher networks, respectively. $|X_k|$ represents the absolute value of each pixel within the channel, and L_{space} is defined as the localized spatial attention feature shared by the student and the teacher. G_{space} represents the aggregate spatial global attention feature of both the student and

the teacher; C is the number of feature channels; n is the number of patches; and T_{space} is categorized as a spatial fusion feature. Following this methodology, we use a similar approach to determine the local and global channel attention features of both the student and teacher networks.

$$B_{S/T}(X) = HW \cdot Softmax\left(\frac{\frac{1}{HW} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} |X_{i,j}|}{\tau}\right) \tag{13}$$

$$P(X) = \left(X^1 \dots X^n\right) \tag{14}$$

$$L_{channel}(P(X_t), P(X_s)) = \varphi\left(B_T\left(X_t{}^1\right) \dots B_T(X_t{}^n)\right) + \varphi\left(B_S\left(X_s{}^1\right) \dots B_S(X_s{}^n)\right) \tag{15}$$

$$G_{channel}(X_t, X_s) = B_T(X_t) + B_S(X_s) \tag{16}$$

$$T_{channel} = \frac{L_{channel}(P(X_t), P(X_s)) + G_{channel}(X_t, X_s)}{2} \tag{17}$$

$B_{S/T}(\cdot)$ is a channel attention operation applied to the student's or teacher's features. $|X_{i,j}|$ denotes the absolute value of each pixel within the spatial dimensions. $L_{channel}$, $G_{channel}$, and $T_{channel}$ each represent the local channel attention features, global channel attention features, and channel fusion features, respectively.

Subsequently, we perform a threshold masking operation on the acquired T_{space} and $T_{channel}$. In this context, a threshold value denoted by θ (θ is set to 1) is used to selectively attenuate redundant information in T_{space} and $T_{channel}$. The threshold filtering formulas are shown below:

$$M_s^{i \subseteq [0,H), j \subseteq [0,W)} = \begin{cases} 0, & T_{space}^{i,j} \geq \theta \\ 1, & T_{space}^{i,j} < \theta \end{cases} \tag{18}$$

$$M_c^{(k \subseteq 0,C)} = \begin{cases} 0, & T_{channel}^{k} \geq \theta \\ 1, & T_{channel}^{k} < \theta \end{cases} \tag{19}$$

where M_s and M_c are spatial threshold mask features and channel threshold mask features, respectively. We use the threshold mask features to code the features of the teacher and student models, respectively, to obtain the channel and spatial mask features, $F_c^{S/T}$ and $F_s^{S/T}$, for the student and teacher.

$$F_s^{S/T} = L_s^{S/T} \otimes M_s \tag{20}$$

$$F_c^{S/T} = L_c^{S/T} \otimes M_c \tag{21}$$

In this context, $L_s^{S/T}$ and $L_c^{S/T}$ represent the spatial and channel feature maps of the teacher and student models, respectively, and $\otimes$ denotes element-wise multiplication.

$$F_{rec}^{S/T} = \frac{Conv\left(F_s^{S/T}\right) + MLP\left(F_c^{S/T}\right)}{2} \tag{22}$$

The spatial and channel mask features of the student and the teacher are input into $Conv()$ and $MLP()$, respectively, for feature reconstruction. The SIMM loss is used to optimize the training process. LCDD offers an innovative and effective solution for target detection in UAV imagery, effectively overcoming the shortcomings of traditional KD methods in dealing with local and semantic information. The LCDD structure is shown in Figure 8.

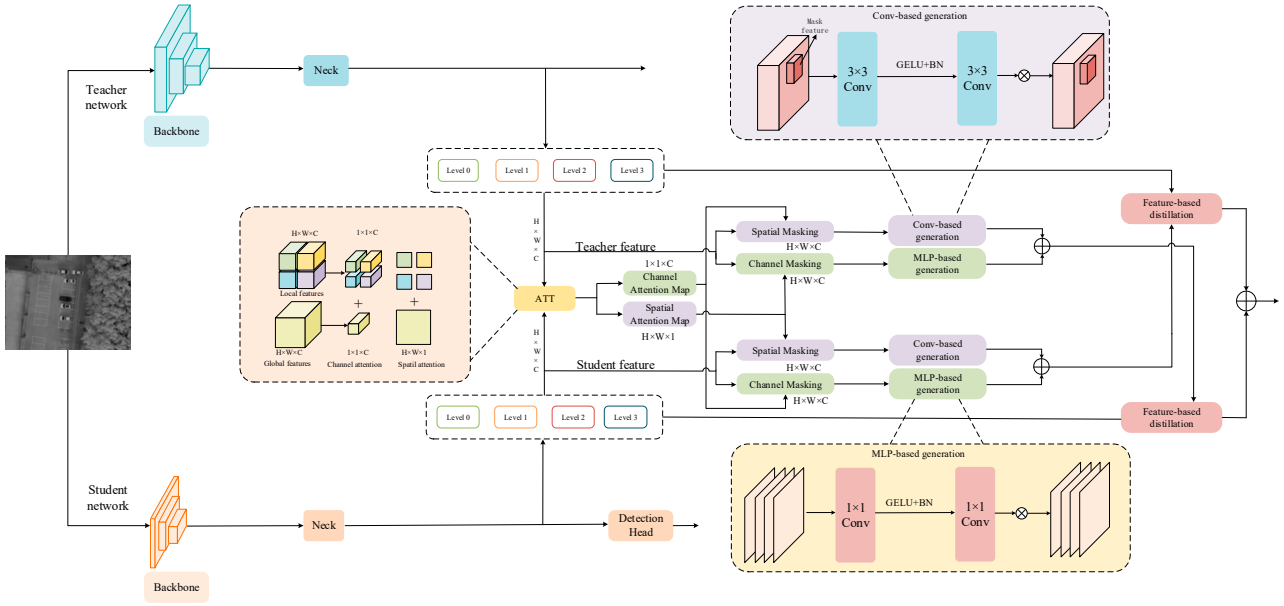

Figure 8. The Structure of the LCDD.

3.6. Loss Function

The loss function is used to measure the distance between the prediction result of the model and the real label. The loss function in this paper is categorized into regression loss and classification loss. Complete Intersection over Union (CioU) loss, denoted as $Loss_{CIoU}$, and Distribute Focal Loss (DFL), denoted as $Loss_{DFL}$, are used to calculate the classification loss. The classification loss was expressed as $Loss_{Cls}$ using the binary cross-entropy loss (BCE). The loss function employed by the proposed MGFAFNET for efficient object detection in UAV imagery is given by the following equation:

$$Loss = \alpha Loss_{CIoU} + \beta Loss_{DFL} + \gamma Loss_{Cls} \tag{23}$$

where $Loss$ represents the total loss and α, β, and γ represent the weights of different losses. The default settings for these weights are 1, 1, and 1.

The $Loss_{CIoU}$ under consideration evaluates the correlation between the predicted value and the actual value in terms of the coordinates of center points, the overlapping area, and the aspect ratio, which is defined as follows:

$$Loss_{CIoU} = 1 - IoU + \frac{d^2}{c^2} + \frac{v^2}{(1 - IoU) + v} \tag{24}$$

$$v = \frac{4}{\pi^2} * \left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w_p}{h_p} \right)^2 \tag{25}$$

where d is the distance between the center points of the prediction box and the real box; c is the diagonal length of the smallest enclosing rectangle that covers both the predicted and actual boxes; v measures the consistency of aspect ratio, comparing the width-to-height ratios of the two boxes.

The BCE loss function is used for confidence loss and classification loss. Its definition is given as follows:

$$Loss_{Cls} = -\left[y_{gt} * log(x_p) + (1 - y_{gt}) * log(1 - x_p) \right] \tag{26}$$

where y_{gt} represents the label confidence and x_p represents the prediction confidence.

4. Experimental Results and Analysis

In this section, we first describe implementation details, the dataset, and evaluation metrics. Subsequently, we compare the performance of our method with existing SOTA on two datasets. Next, we conducted ablation experiments to analyze the contribution of different modules. To explore the effectiveness of each module in more depth, we provide more details module analysis results. Finally, we performed a visual analysis of the results to demonstrate the superiority of our approach.

4.1. Implementation Details

The experimental server uses an Intel Core2 i7-6950X CPU processor and NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory. All the methods were implemented on a Linux operating system (Ubuntu 18.04.1) and trained offline based on a Pytorch1.6.0 deep learning framework. Inference tests were performed on a NVIDIA XAVIER NX and NVIDIA RTX-4090.The NVIDIA XAVIER NX weighs 200 g and has a volume of 7 cm $\times$ 5 cm $\times$ 4 cm, which meets the requirements for UAV applications. The stochastic gradient descent (SGD) was used to optimize our network. A total of 300 periods were trained. Weight emphasis and momentum were 0.0005 and 0.937. The batch size was set to be 32. The initial learning rate was 0.01, and it was reduced to 0.1 every 100 epochs of training. In addition, various data-enhancement schemes, such as Mosaic and Mixup, were used for image.

In the KD experiment, we followed the above steps to train the MGFAFNET-S model as the teacher model, while the MGFAFNET-N model was used as the student model. We used the pre-trained MGFAFNET-S teacher model for network training for a total of 300 epochs. The temperature parameter τ was set to 1, while the mask threshold was set to 1. Other parameters were set to the same values as used for normal training.

4.2. Dataset Preparation

(1) SyluDrone dataset: We constructed the SyluDrone dataset with the help of an independently designed UAV platform. To ensure the authenticity of the dataset, we used the UAV to capture images at different heights and pitch angles. The SyluDrone dataset covers 6534 real images in a variety of scenarios such as urban and suburban areas, each with a pixel resolution of 1920 $\times$ 1280.We annotated the target bounding boxes in the images by manually labeling them. In our study, we focused on the category "car". This dataset provides a valuable resource for studying the field of UAV vision, especially in object detection.

(2) The VisDrone2021 dataset is a comprehensive UAV image repository designed for object detection, single-object tracking, and multi-object tracking tasks. It features a wide variety of categories, including pedestrians, people, cars, vans, buses, trucks, motorcycles, bicycles, awning tricycles, and tricycles, totaling 10,209 still images. Each image is approximately 2000 $\times$ 1500 pixels in resolution. The dataset is divided into three parts: a training set with 6471 images, a validation set with 548 images, and a test set comprising 3190 images. All images in the VisDrone2021 were captured using various UAV platforms across diverse lighting and weather conditions.

The proportions of goals of different sizes in the two datasets are shown in Figure 9. As can be seen from the figure, the largest proportion of small objects is represented.

4.3. Evaluation Metrics

In order to qualitatively analyze the performance of the proposed model, we use the most common evaluation metric map in object-detection tasks to evaluate the model: the mean average precision (mAP). mAP calculation involves taking different intersection-over-union (IoU) thresholds from 0.5 to 0.95 and a step size of 0.05. mAP is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{27}$$

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

$$mAP = \frac{1}{n}\sum_{k=1}^{n} J(Precision, Recall)_k \tag{29}$$

where TP represents the true-positive real examples, FP represents the false-positive examples, FN represents the false-negative examples, n represents the number of categories, and $J(Precision, Recall)_k$ represents the average precision function. We use frames per second (FPS) to measure the detection speed, which represents the number of images the detection model can process per second with the specified hardware. The detection rate is controlled by the IoU between the detection result and ground truth. If the IoU is greater than the threshold, the result is considered a true detection.

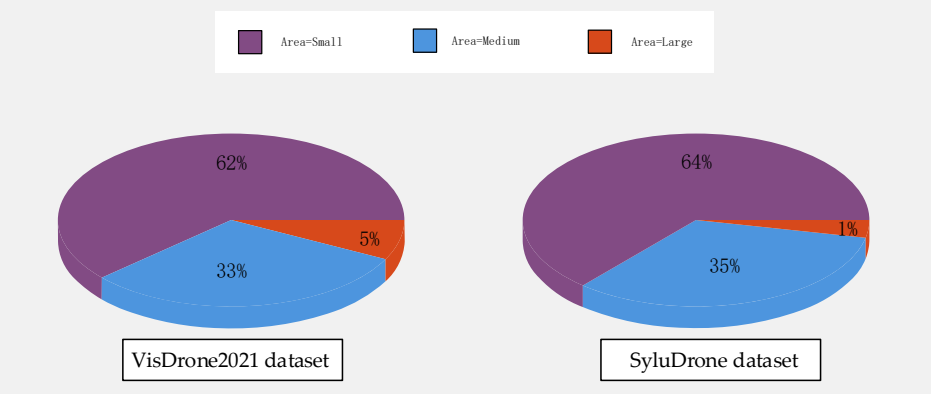

Figure 9. Distribution of Object Sizes Across Two Datasets.

4.4. Comparisons with SOTA Method

To comprehensively evaluate the superiority of the proposed method, we evaluate MGFAFNET against current SOTA methods. Specifically, the comparison methods are divided into two groups, the first group includes one-stage detectors, including YOLOv5, YOLOV6n [34], YOLOV7-tiny [8], and YOLOv8n.The second group comprises two-stage detectors, including Cascade-RCNN [35], Light-RCNN [36], and FasterR-CNN [37]. From the experimental results, it can be observed that the proposed model achieves the best results on AP.

Results from the VisDrone2021 datasets are shown in Table 1. For the two-stage detectors, the MGFAFNET-N significantly outperforms Cascade-RCNN, Light-RCNN, and Faster R-CNN with higher complexity by 3.7%, 2.8%, and 12.4%, respectively. For the one-stage detectors, our proposed MGFAFNET significantly outperforms the corresponding networks from the YOLOv5 and YOLOv8 series, while maintaining comparable complexity. Specifically, MGFAFNET-N realizes a performance increase of 8.3% over YOLOv5n, and MGFAFNET-S surpasses YOLOv5s by 3.0%. Similarly, when comparing MGFAFNET models to those in the YOLOv8 series, we observe a consistent improvement: MGFAFNET-N outstrips YOLOv8n by 6.6%; MGFAFNET-S outperforms YOLOv8s by 3.5%; and MGFAFNET-M outperforms YOLOv8m by 2%. In addition, compared to other models in the YOLO series, MGFAFNET-N offers a 10.6% improvement over YOLOv6n and a 5% improvement over YOLOv7-tiny. Our MGFAFNET exhibits relatively better detection performance while having fewer parameters. For example, at similar or higher accuracies, MGFAFNET-N has 5M, 2.6M, 4.1M, and 9.1M fewer parameters than the representative one-stage detectors YOLOv5s, YOLOv6n, YOLOv7-tiny, and YOLOv8s. It is worth noting that all these representative one-stage detectors are very lightweight, while our MGFAFNET-N is smaller than these lightweight one-stage detectors.

Table 1. Comparison of experimental results on Visdrone2021 datasets.

Method	Input Size	Params (M)	GFLOPs	AP (%)	AP_{50} (%)	FPS
Cascade-RCNN	1536 × 1536	—	—	16.1	31.9	—
Light-RCNN	1536 × 1536	—	—	16.5	32.8	—
FasterR-CNN	1536 × 1536	—	—	12.1	23.5	—
YOLOv5n	640 × 640	1.8	4.2	14.2	27.3	158.6
YOLOV5s	640 × 640	7.0	15.8	18.6	33.6	130.6
YOLOV6n	640 × 640	4.6	11.34	14.0	25.0	211.0
YOLOV7-tiny	640 × 640	6.1	13.1	15.8	30.6	243.0
YOLOv8n	640 × 640	3.0	8.1	16.4	29.0	240.4
YOLOv8s	640 × 640	11.1	28.5	19.2	33.1	208.9
YOLOv8m	640 × 640	25.8	78.7	21.5	36.8	102.8
MGFAFNET-N	640 × 640	2.0	13.5	20.2	35.6	141.1
MGFAFNET-S	640 × 640	4.2	24.2	21.0	36.6	116.9
MGFAFNET-M	640 × 640	7.5	39.7	22.2	38.8	82.6

Results from the SyluDrone datasets are shown in Table 2. For the one-stage detectors, MGFAFNET-N significantly outperforms YOLOv5n, YOLOv6n, YOLOv7-tiny, and YOLOv8n by 10.6%, 7.4%, 2.4%, and 5.2%, respectively. MGFAFNET-S exceeds YOLOv5s and YOLOv8s by 9.1% and 6.8%, respectively. The MGFAFNET-M exceeds YOLOv8m by 6.8%. The experimental results fully demonstrate that our method can handle complex backgrounds and small objects in UAV images more efficiently and show that the method has significant advantages in terms of the number of parameters and AP_{50}.

Table 2. Comparison of experimental results from SyluDrone datasets.

Method	Input Size	Params (M)	GFLOPs	AP (%)	AP_{50} (%)	FPS
YOLOv5n	640 × 640	1.77	4.1	31.0	79.6	158.1
YOLOV5s	640 × 640	7.03	15.8	34.3	83.3	130.2
YOLOV6n	640 × 640	4.63	11.34	31.6	82.8	210.6
YOLOV7-tiny	640 × 640	6.05	13.0	45.6	87.8	242.0
YOLOv8n	640 × 640	3.00	8.1	45.4	85.0	239.9
YOLOv8s	640 × 640	11.12	28.4	46.9	86.1	207.1
YOLOv8m	640 × 640	25.84	78.7	48.3	86.8	103.1
MGFAFNET-N	640 × 640	2.09	13.5	50.1	90.2	141.2
MGFAFNET-S	640 × 640	4.28	24.2	51.6	92.4	116.6
MGFAFNET-M	640 × 640	7.56	39.7	52.7	93.6	82.4

4.5. Ablation Study

To evaluate the contribution of our three proposed modules (DBMA backbone, DSPF neck, and LCDD distillation) to the performance improvement of MGFAFNET model, we conducted four sets of ablation experiments to investigate different network architectures using the VisDrone2021 dataset and our SyluDrone dataset; the experimental results are shown in Table 3. We selected YOLOv8n as the baseline model.

With the VisDrone2021 dataset and our SyluDrone dataset, the baseline model achieved AP_{50} scores of 29.0% and 85.0%, respectively. However, our model demonstrated optimal performance with the least parameters, achieving AP_{50} scores of 35.6% and 90.2%, respectively. All proposed modules exhibited improvements in detection accuracy. Notably, the DBMA's accuracy in extracting complex background features increased by 2.4% and 1.8%, respectively, while reducing the number of parameters by 0.05 million. This result emphasizes DBMA's ability to efficiently handle objects in complex scenes in a lightweight manner. Similarly, when the DSPF is utilized to extract multi-scale features, the AP_{50} scores improved by 2% and 2.9%, respectively, while the number of parameters was reduced by 0.83 million. In addition, DSPF performs better for small objects with AP_S scores improved by 1.8% and 3.7%, respectively. This indicates that DSPF is more adept at managing

multi-scale variations in UAV images, especially for small objects. When utilizing both DBMA and DSPF, the AP_{50} scores were improved by 5.5% and 3.4%, respectively, while the number of parameters was reduced by 0.91 million and the AP_S scores improved by 3.1% and 4%, respectively.

Table 3. The results of ablation experiments on two datasets.

Visdrone Dataset													
Baseline	**DBMA**	**DSPF**	**LCDD**	**Params (M)**	**GFLOPs**	**AP (%)**	**AP_{50} (%)**	**AP_S (%)**	**AP_M (%)**	**AP_L (%)**	**FPS (GPU)**	**FPS (Xavier nx)**	**Memory (G)**
√				3.00	8.1	16.4	29.0	6.1	23.8	34.8	241.0	35.3	1.6
√	√			2.95	8.9	17.6	31.4	7.2	25.6	38.1	183.2	30.6	1.6
√		√		2.17	12.7	17.5	31.0	7.9	24.4	35	164.2	28.4	1.6
√	√	√		2.09	13.5	19.4	34.5	9.2	26.8	34.6	143.5	27.2	1.6
√	√	√	√	2.09	13.5	20.2	35.6	9.7	28.6	36.1	143.5	27.2	1.7
SyluDrone Dataset													
Baseline	**DBMA**	**DSPF**	**LCDD**	**Params (M)**	**GFLOPs**	**AP (%)**	**AP_{50} (%)**	**AP_S (%)**	**AP_M (%)**	**AP_L (%)**	**FPS (GPU)**	**FPS (Xavier nx)**	**Memory (G)**
√				3.00	8.1	45.4	85.0	34.1	55.2	55.9	241.0	35.2	1.6
√	√			2.95	8.9	46.6	86.8	36.5	55.9	47.5	182.2	30.4	1.6
√		√		2.17	12.7	47.6	87.9	37.8	57.7	51.6	162.8	28.4	1.6
√	√	√		2.09	13.5	48.8	88.4	38.1	58.8	53.1	144.1	27.1	1.6
√	√	√	√	2.09	13.5	50.1	90.2	38.9	60.4	56.8	144.1	27.1	1.7

The results from the ablation experiments involving two modules indicate that the combined use of both DBMA and DSPF not only results in a reduction of parameters but also achieves enhanced performance compared to the isolated application of each module. When the DBMA backbone, DSPF neck, and LCDD distillation methods were implemented jointly, the AP_{50} scores improved by 6.6% and 5.2%, respectively. For small objects, the AP_S scores also increased by 3.6% and 4.8%. For medium-sized objects, the AP_M scores increased by 4.8% and 5.2%, and for large objects, the AP_L scores increased by 1.3% and 0.9%, respectively. In addition, the experimental results show that the LCDD refinement technique significantly improves the accuracy of the baseline model without introducing additional parameters or computational overhead. Despite a marginal increase in computational demand, the system remains compliant with the real-time requirements of UAV operations.

The results of the ablation experiments clearly confirm the effectiveness of each proposed module and of the combination of different modules.

4.6. Module Analysis

In this section, we perform a comprehensive model analysis of the proposed two key modules to better illustrate the effectiveness of our proposed framework. We compare the DBMA backbone with several SOAT methods, including Efficientformerv2-s0 [38], RepViT-M1 [39], Fasternet-T0 [40], and ConvNextv2-A [41].

Efficientformerv2-s0 leverages a fine-grained joint search algorithm to achieve low latency and high parameter efficiency. RepViT combines the architectural choices of lightweight ViTs and was specifically designed for resource-constrained mobile devices. Fasternet-T0 enhances computation speed on mobile devices by reducing redundant calculations and memory access. Convnextv2-A extends the ConvNeXt architecture, evolving into a fully convolutional mask autoencoder backbone network. It is worth noting that we excluded LCDD in order to make it easier to assess the effectiveness of DBMA and DSPF when performing the modeling analysis on trunks and necks.

The experimental results from the Visdrone2021 dataset are shown in Figure 10a. As the smallest model among the five backbones in terms of dimensions, the DBMA trunks still achieved highly competitive detection accuracy with an AP_{50} of 34.5%, and the computational complexity is significantly lower than that of existing SOAT methods. Specifically, DBMA surpasses the second-best RepViT-M1 by 0.4% in AP_{50} with 2.38M fewer parameters and surpasses the model with Efficientformerv2-s0 backbone by 1.7% in

AP_{50} while using about half the number of parameters. DBMA outperforms Fasternet-T0 by 3.5% in AP_{50} with 1.28M fewer parameters and Convnextv2-A by 5.2% in AP_{50} with 2.78M fewer parameters. In terms of computational efficiency, DBMA has also demonstrated a leading advantage. Compared to Efficientformerv2-s0, RepViT-M1, Fasternet-T0, and Convnextv2-A, DBMA achieves reductions in computational complexity by 2.8%, 9.9%, 2%, and 5.4%, respectively. The experimental results validate the ability of the DBMA backbone to efficiently capture both local and long-range dependencies.

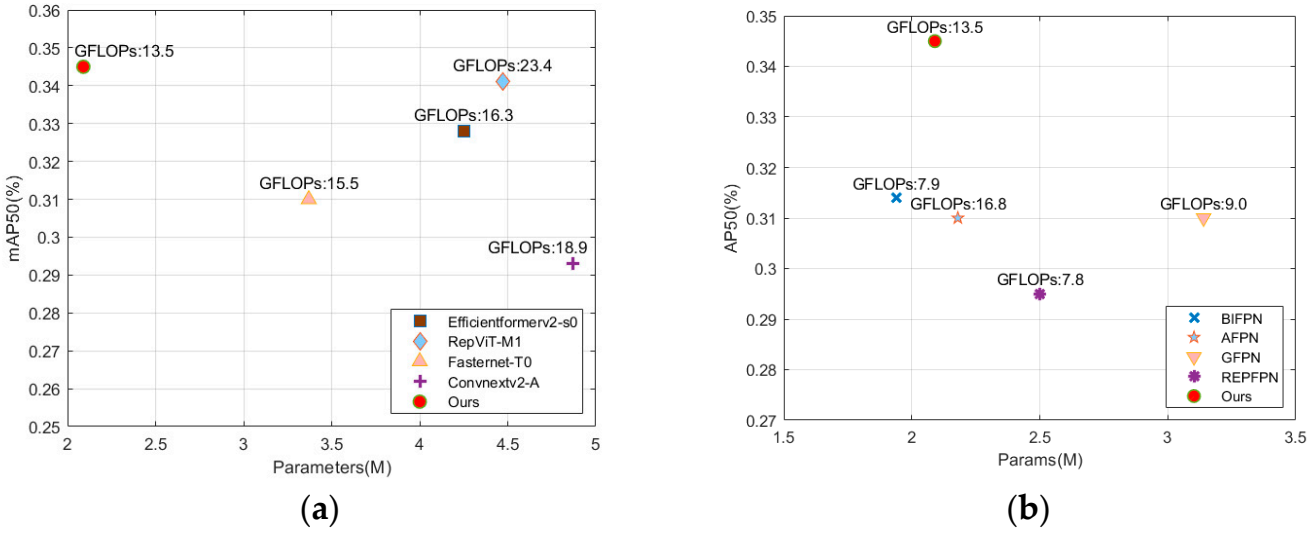

Figure 10. Experimental results of backbone and neck modeling analyses. (**a**) Analysis of different backbone Models; (**b**) Analysis of Different Neck Models.

In addition, we compared the DSPF neck with four SOAT methods (BiFPN [42], AFPN [43], GFPN [44] and Rep-FPN [8]) on the Visdrone2021 dataset. BiFPN effectively optimize multi-scale feature fusion by introducing bidirectional connectivity and additional fusion weights, AFPN effectively prevents the loss of feature information during transmission and interaction by facilitating direct feature fusion between nonadjacent layers. GFPN processes both high-level semantic information and low-level spatial information at the same time with equal priority, effectively integrating multi-scale complementary features. Rep-FPN utilizes the ideas of structural reparameterization and hardware-aware neural network design to effectively improve the accuracy of object detection.

The experimental results are shown in Figure 10b. As can be seen from the results, our method obtained the highest AP_{50}, at 34.5%. As for accuracy, with comparable or lower parameters, DSPF outperforms BiFPN, AFPN, GFPN, and Rep-FPN by 3.1%, 3.5%, 3.5% and 5.0% AP_{50}, respectively. Compared with the second-best method, BiFPN, the DSPF improves accuracy by 3.1%. The experiments better demonstrate that DSPF provides more detailed and precise control of information flow and fusion when dealing with the challenge of multi-scale detection in aerial images and enhances the model's adaptability to multi-scale scenes.

4.7. Analysis of Visualization Results

To further demonstrate the effectiveness of the proposed MGFAFNET, we show in Figure 11a the visual detection results for common problems related to object detection in UAV images (multi-scale objects, complex backgrounds, object occlusion, and small objects) in the Visdrone2021 dataset.

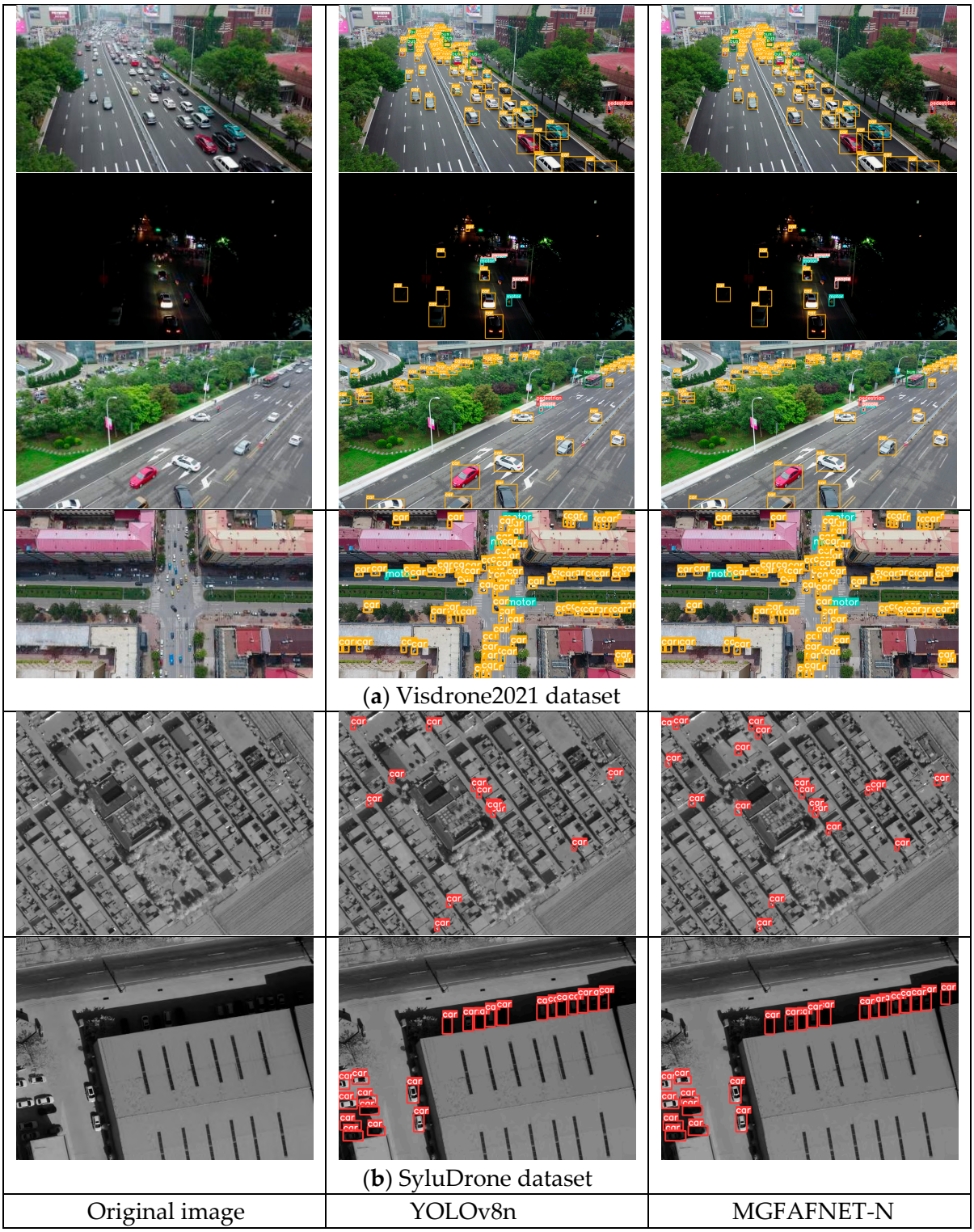

Figure 11. Detection of visualization results in the Visdrone2021 dataset and the SyluDrone dataset.

The first column of the figure presents the original image, while the second and third columns show the prediction results of the YOLOv8n and MGFAFNET models, respectively. The diversity of object scales is a common problem in UAV image detection, and as can be seen from the visualization results in the first row, our proposed MGFAFNET shows significant advantages over Yolov8n in handling multi-scale objects more effectively. Although yolov8n detects most of the objects in a dark environment, it is still less effective

in performing the complex background task. By contrast, MGFAFNET is not affected by these factors and detects more targets in the dark than does the baseline model yolov8n.In addition, the third row of results reveals a remarkable phenomenon: the discontinuity of target information due to the presence of occluded objects leads to a significant degradation in YOLOv8n's performance in detecting a large number of small objects. Finally, to verify the effectiveness of the proposed method in small-object detection, we visualize the small-target detection results. As can be seen from the fourth row of results, MGFAFNET is able to accurately determine the true contours of dense small objects and detects more small targets than does the yolov8n.

In the visualization results for the SyluDrone dataset, as depicted in Figure 11b, we present the detection outcomes of our method. It is evident from the results in the fifth row that compared to the baseline model yolov8n, our approach can more effectively detect a greater number of small targets. This finding highlights the superiority of our method in object-detection tasks, indicating its enhanced adaptability and robustness. Upon further examination of the results in the sixth row, we observe that our method performs exceptionally well in detecting vehicles against complex backgrounds, achieving complete detection of targets. This finding implies that our approach can deliver satisfactory detection performance in challenging environments, thereby offering a reliable solution for target-detection tasks in drone application scenarios.

5. Conclusions

In this article, we introduce a groundbreaking object-detection framework tailored to UAV-based applications, which we called the Efficient Multidimensional Global Feature Adaptive Fusion Network (MGFAFNET). This innovative network distinguishes itself by adeptly managing the delicate equilibrium between computational efficiency and detection performance in UAV optical sensor-captured conditions. The MGFAFNET is underpinned by three main innovative components: the Dual-Branch Multidimensional Aggregation Backbone Network (DBMA), which revolutionizes the encoding of global features by efficiently overcoming the computational intensity often associated with self-attention token mixers in UAV imagery; the Dynamic Spatial Perception Feature Fusion Network (DSPF), which represents a significant leap forward in suppressing extraneous features through multi-level adaptive fusion, thereby bolstering the network's ability to perform multi-scale detection within the varied UAV imaging domainl and the Localized Compensation Dual-Mask Distillation (LCDD), which brings a novel strategy to the table by harnessing both local and global fusion features from both teacher and student networks to enhance feature distillation, resulting in unprecedented detection precision in UAV images. Moreover, we have constructed an advanced UAV image object-detection platform to rigorously test the proposed MGFAFNET. Remarkable results from our extensive experiments on the self-built dataset underscore MGFAFNET's superiority, with an impressive 90.2% AP_{50} and demonstrably outperformance of the current state-of-the-art (SOTA) methods. Complementing this finding, MGFAFNET has also shown its versatility in application to the public VisDrone2021 dataset, with a substantial 35.6% AP_{50}. These findings underscore the robust generalization potential of MGFAFNET to meet the intricate demands of UAV detection across a spectrum of complex scenarios, firmly establishing our method as a significant stride forward in UAV-based object-detection technology. Compared to SOAT detection algorithms, our method exhibits slightly higher computational complexity, necessitating future research aimed at reducing computational overhead to improve practicality and efficiency while maintaining performance.

Author Contributions: Conceptualization, J.S., J.Y. and H.G.; methodology, J.S. and Z.Y.; software, Z.J.; validation, Z.J.; for-mal analysis, J.S.; investigation, H.G.; resources, H.G.; data curation, Z.Y.; writing—original draft preparation, J.S.; writing—review and editing, J.Y.; visualization, X.Q.; supervision, J.Y.; project administration, Z.Y.; funding acquisition, H.G. All authors have read and agreed to the published version of the manuscript.

Funding: Technology Development Program (JCKY2022410C002). LiaoNing Province Joint Open Fund for Key Scientific and Technological Innovation Bases (Grant No.2021-KF-12-05).0pen fund for the State Key Laboratory of Robotics (2023003).

Data Availability Statement: Data are contained within the article.

Acknowledgments: Thanks to Shenyang Ligong University for providing the equipment and laboratory.

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

References

1. Ke, R.; Li, Z.; Tang, J.; Pan, Z.; Wang, Y. Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 54–64. [CrossRef]
2. Su, J.; Zhu, X.; Li, S.; Chen, W.-H. AI meets UAVs: A survey on AI empowered UAV perception systems for precision agriculture. *Neurocomputing* **2023**, *518*, 242–270. [CrossRef]
3. Mittal, P.; Singh, R.; Sharma, A. Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image Vis. Comput.* **2020**, *104*, 104046. [CrossRef]
4. Yu, J.; Gao, H.; Chen, Y.; Zhou, D.; Liu, J.; Ju, Z. Adaptive spatiotemporal representation learning for skeleton-based human action recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 1654–1665. [CrossRef]
5. Yu, J.; Gao, H.; Zhou, D.; Liu, J.; Gao, Q.; Ju, Z. Deep temporal model-based identity-aware hand detection for space human–robot interaction. *IEEE Trans. Cybern.* **2021**, *52*, 13738–13751. [CrossRef]
6. Yu, J.; Gao, H.; Chen, Y.; Zhou, D.; Liu, J.; Ju, Z. Deep object detector with attentional spatiotemporal LSTM for space human–robot interaction. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 784–793. [CrossRef]
7. Girshick, R. Fast r-cnn. In Proceedings of the IEEE international Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
9. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
10. Yu, J.; Xu, Y.; Chen, H.; Ju, Z. Versatile Graph Neural Networks Toward Intuitive Human Activity Understanding. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–13. [CrossRef]
11. Yu, J.; Zheng, W.; Chen, Y.; Zhang, Y.; Huang, R. Surrounding-aware representation prediction in Birds-Eye-View using transformers. *Front. Neurosci.* **2023**, *17*, 1219363. [CrossRef]
12. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.
13. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive Sparse Convolutional Networks with Global Context Enhancement for Faster Object Detection on Drone Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13435–13444.
14. Zhou, W.; Min, X.; Hu, R.; Long, Y.; Luo, H. Faster-X: Real-Time Object Detection Based on Edge GPUs for UAV Applications. *arXiv* **2022**, arXiv:2209.03157.
15. Lu, W.; Lan, C.; Niu, C.; Liu, W.; Lyu, L.; Shi, Q.; Wang, S. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211–1231. [CrossRef]
16. Yang, G.; Tang, Y.; Wu, Z.; Li, J.; Xu, J.; Wan, X. DMKD: Improving Feature-based Knowledge Distillation for Object Detection Via Dual Masking Augmentation. *arXiv* **2023**, arXiv:2309.02719.
17. Jang, Y.; Shin, W.; Kim, J.; Woo, S.; Bae, S.H. GLAMD: Global and Local Attention Mask Distillation for Object Detectors. In *European Conference on Computer Vision*; Springer Nature Switzerland: Cham, Switzerland, 2022; pp. 460–476.
18. Yue, K.; Deng, J.; Zhou, F. Matching guided distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XV 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 312–328.
19. Yang, G.; Tang, Y.; Li, J.; Xu, J.; Wan, X. AMD: Adaptive Masked Distillation for Object Detection. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
20. Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 284–2854.
21. Ye, T.; Qin, W.; Li, Y.; Wang, S.; Zhang, J.; Zhao, Z. Dense and small object detection in UAV-vision based on a global-local feature enhanced network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13. [CrossRef]

22. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient Detection of UAV Image Based on Cross-layer Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [CrossRef]
23. Liao, J.; Piao, Y.; Su, J.; Cai, G.; Huang, X.; Chen, L.; Huang, Z.; Wu, Y. Unsupervised cluster guided object detection in aerial images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11204–11216. [CrossRef]
24. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A global-local self-adaptive network for drone-view object detection. *IEEE Trans. Image Process.* **2020**, *30*, 1556–1569. [CrossRef]
25. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
26. Li, W.; Wang, X.; Xia, X.; Wu, J.; Xiao, X.; Zheng, M.; Wen, S. Sepvit: Separable vision transformer. *arXiv* **2022**, arXiv:2203.15380.
27. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
28. Wang, J.; Chen, Y.; Zheng, Z.; Li, X.; Cheng, M.M.; Hou, Q. CrossKD: Cross-Head Knowledge Distillation for Dense Object Detection. *arXiv* **2023**, arXiv:2306.11369.
29. Yang, L.; Zhou, X.; Li, X.; Qiao, L.; Li, Z.; Yang, Z.; Wang, G.; Li, X. Bridging Cross-task Protocol Inconsistency for Distillation in Dense Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 17175–17184.
30. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. *arXiv* **2014**, arXiv:1412.6550.
31. Zhang, L.; Ma, K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In Proceedings of the International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
32. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
33. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
34. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
35. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
36. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Light-head r-cnn: In defense of two-stage object detector. *arXiv* **2017**, arXiv:1711.07264.
37. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
38. Li, Y.; Hu, J.; Wen, Y.; Evangelidis, G.; Salahi, K.; Wang, Y.; Tulyakov, S.; Ren, J. Rethinking vision transformers for mobilenet size and speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 16889–16900.
39. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. Repvit: Revisiting mobile cnn from vit perspective. *arXiv* **2023**, arXiv:2307.09283.
40. Chen, J.; Kao, S.H.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12021–12031.
41. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142.
42. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
43. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. Afpn: Asymptotic feature pyramid network for object detection. *arXiv* **2023**, arXiv:2306.15988.
44. Jiang, Y.; Tan, Z.; Wang, J.; Sun, X.; Lin, M.; Li, H. GiraffeDet: A heavy-neck paradigm for object detection. *arXiv* **2022**, arXiv:2202.04256.