A Lightweight Fusion Strategy with Enhanced Inter-layer Feature Correlation for Small Object Detection

Yao Xiao, Tingfa Xu*, Xin Yu , Yuqiang Fang and Jianan Li*

***Abstract*—Detecting small objects in drone imagery is challenging due to low resolution and background blending, leading to limited feature information. Multi-scale feature fusion can enhance detection by capturing information at different scales, but traditional strategies fall short. Simple concatenation or addition operations do not fully utilize multi-scale fusion advantages, resulting in insufficient correlation between features. This inadequacy hinders the detection of small objects, especially in complex backgrounds and densely populated areas. To address this issue and efficiently utilize the limited computational resources, we propose a lightweight fusion strategy based on enhanced inter-layer feature correlation (EFC) to replace the traditional feature fusion strategy in FPN. The semantic expressions of different layers in the feature pyramid are inconsistent. In EFC, the Grouped Feature Focus Unit (GFF) enhances the feature correlation of each layer by focusing on the contextual information of different features. The Multi-Level Feature Reconstruction Module (MFR) effectively reconstructs and transforms the strength and weakness information of each layer in the pyramid to reduce redundant feature fusion and retain more information about small targets in deep networks. It is noteworthy that the proposed method is plug-and-play and can be widely applied to various base networks. Extensive experiments and comprehensive evaluations on VisDrone, UAVDT and COCO demonstrate the effectiveness. Using GFL as the baseline on the VisDrone dataset with a large number of small targets, the proposed method improves the detection mAP by 1.7%, surpassing many lightweight state-of-the-art methods and significantly reducing the Params and GFLOPs at the neck end.**

***Index Terms*—Small object detection, feature fusion, lightweight**

I. Introduction

In recent years, object detection for drones has gained significant momentum in applications such as remote sensing, traffic monitoring, search operations, and security [1], [2], [3], [4], [5]. The successful implementation of these applications relies on the rapid and effective recognition of objects within drone-captured images. Drone object detection presents unique challenges. On the one hand, drone images are characterized by high resolution and significant background noise, which can obscure target objects and make detection more difficult. On the other hand, the hardware limitations of drones necessitate improvements in detection accuracy within these resource constraints.

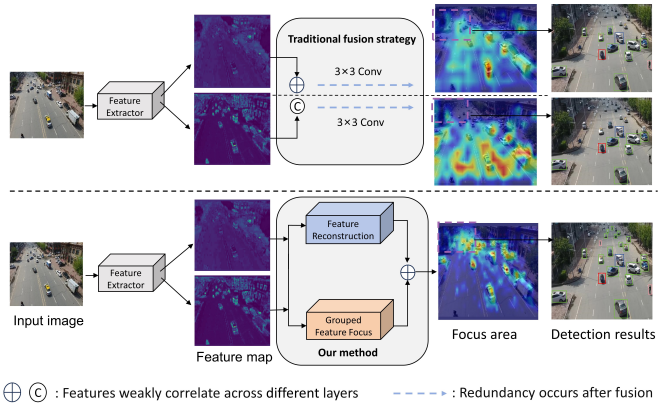

Fig. 1. The traditional fusion strategy has several drawbacks, including weak correlation between features from different layers and the generation of redundant features after fusion. Detecting small objects is challenging due to their small size, which makes them prone to information loss in deep networks and susceptible to being overshadowed by background noise. Grouped feature focus unit (GFF) significantly improves the representation of small object information across feature maps, while feature reconstruction (MFR) addresses the issue of irreversible loss of small object information in deep networks.

Small object detection, in particular, is a challenging task in drone imagery due to the low resolution of small objects, which makes them more susceptible to noise and results in limited effective information. During convolutional neural network (CNN) feature extraction, small objects are prone to feature disappearance. To address this, multi-scale feature fusion enhances the network's ability to perceive small objects by enriching their feature representation. This approach captures target information at various scales and provides high-resolution feature maps, which are essential for precise localization and recognition.

The Feature Pyramid Network (FPN) [6] is a widely adopted model that generates pyramid-style multi-scale features by integrating adjacent layers in the backbone network. This fusion combines deep abstract semantic features with shallow high-resolution features, thereby enhancing feature expression across different scales. However, traditional fusion methods in FPN, such as simple concatenation or addition operations, do

*Corresponding author

Yao Xiao, Tingfa Xu, Xin Yu and Jianan Li are with Beijing Institute of Technology, Beijing 100081, China (e-mail:{3220220586, ciom_xtf1,3220220587,lijianan}@bit.edu.cn).

Yuqiang Fang is with the Science and Technology on Complex Electronic System Simulation Laboratory, Space Engineering University, Beijing 101416, China (e-mail: fangyuqiang@nudt.edu.cn).

Jianan Li and Tingfa Xu are also with the Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education of China, Beijing 100081, China.

Tingfa Xu is also with Beijing Institute of Technology Chongqing Innovation Center, Chongqing, 401135, China.

not fully exploit the benefits of multi-scale fusion. As shown in Fig. 1, low-resolution features are upsampled and fused with adjacent layers solely by stacking and channel fusion, without fully considering their correlation. This oversight restricts the utilization of correlated features from each layer, thus diminishing the representation of multi-scale information.

At the feature fusion stage, the fused information is sent to the detection head structure through a 3×3 convolution layer. Utilizing large kernel convolutions in the deep layers often generates redundant information, wasting computational resources. Additionally, features from the top of the pyramid already possess high-level semantic information. Passing them through large kernel convolutions can cause semantic deviations with shallow features, making small target feature information more abstract and difficult to match with their spatial positions.

We note that traditional fusion methods in the neck of the network are insufficient for fully leveraging the benefits of multi-scale feature fusion. It is crucial to perform hierarchical deconstruction of features. In response to the aforementioned issues, we propose a lightweight feature fusion strategy, named EFC (Enhanced Inter-layer Feature Correlation). This strategy extends the conventional two-layer feature map fusion architecture through two specially designed modules. By focusing on spatial contextual information and the commonality between inter-layer features, this strategy enhances semantic representation between features to improve the learning of multi-scale features. Specifically, we first design the Grouped Feature Focusing unit (GFF) to obtain global information and enhance the correlation of fused features. This process mainly includes spatial focusing, feature grouping and fusion, and spatial mapping normalization. This addresses the issues of poor feature correlation and matching in traditional fusion strategies, providing richer contextual information that helps the model more accurately locate small objects. Subsequently, we introduce the Multi-level Feature Reconstruction (MFR) module to replace the 3×3 convolutions in the network's neck. The MFR separates the strong and weak information among features, directs their transformation, and achieves high-level feature aggregation. By reconstructing features, it reduces the semantic deviation that occurs when combining shallow and deep features. This approach aims to reduce feature redundancy introduced by fusion and transformation, minimize the loss of small object feature information in deep networks, and enhance the representation of small objects. It is worth noting that the method can be flexibly applied to various detectors utilizing multi-scale feature fusion.

We validate the effectiveness of our proposed method using mainstream frameworks and various backbone networks on widely used datasets including VisDrone [7], UAVDT [8], and COCO [9]. Experiments demonstrate that our method is effective not only for small object detection but also significantly reduces the parameter count and GFLOPs at the neck stage.

Before delving into the specifics of our method, we summarize our contributions as follows:

- We propose a lightweight feature fusion strategy with enhanced inter-layer feature correlation, termed EFC, which efficiently detects small objects.
- We introduce Grouped Feature Focusing units (GFF) to aggregate contextual information and enhance the correlation between features at different layers.
- We introduce a Feature Reconstruction Module (FRM) to replace large-kernel convolutions in deep networks, reducing the loss of subtle information and the generation of redundant features, thereby minimizing resource consumption.

II. RELATED WORK

A. General Object Detection

Object detection is a fundamental task in computer vision, aimed at identifying the categories and locations of objects within a given image. Currently, object detection methods can be broadly categorized into two types: two-stage detectors based on the RCNN series [10], [11], [12], and single-stage detectors based on the SSD [13] and YOLO series [14]. Two-stage detectors first localize candidate regions in the image and then classify each candidate region using a classifier. On the other hand, single-stage detectors adopt an end-to-end approach, simultaneously performing localization and classification within the same network. Generally, two-stage detection methods tend to achieve higher precision but slower detection speeds compared to single-stage detection methods. However, RetinaNet [15], as a single-stage detector, achieves detection performance on par with two-stage detectors. Similar to GFL [16], it treats anchors as the final bounding box targets. Although these methods perform well in general object detection tasks, they are not satisfactory when applied to drone images that contain a large number of small objects. In this paper, we introduce our proposed small object detection method and apply it to different detectors to demonstrate its effectiveness.

B. Small Object Detection

Small object detection has always been a challenging problem due to the difficulty in localizing objects with limited size. In recent years, many approaches have been proposed to address this issue [17], [18], [19]. Data augmentation techniques have been employed to mitigate the scarcity of small object data [20]. Increasing the resolution of input features enlarges the size of small objects [21], enhancing localization capabilities. However, this significantly increases the model's complexity and reduces detection speed. QueryDet [22] utilizes high-resolution features and introduces a novel query mechanism to accelerate the inference speed of feature pyramid-based object detectors, addressing the issue of high computational resource consumption. CEASC [23] develops a context-enhanced sparse convolution to capture global information and enhance focus features, balancing detection accuracy and efficiency while greatly reducing the computational complexity of detecting high-resolution images.

Most of these methods focus on improving the detection head to enhance network performance. In contrast, our work primarily improves the neck stage, addressing the shortcomings in fusing high-resolution and low-resolution features to solve the problems of weak feature representation and

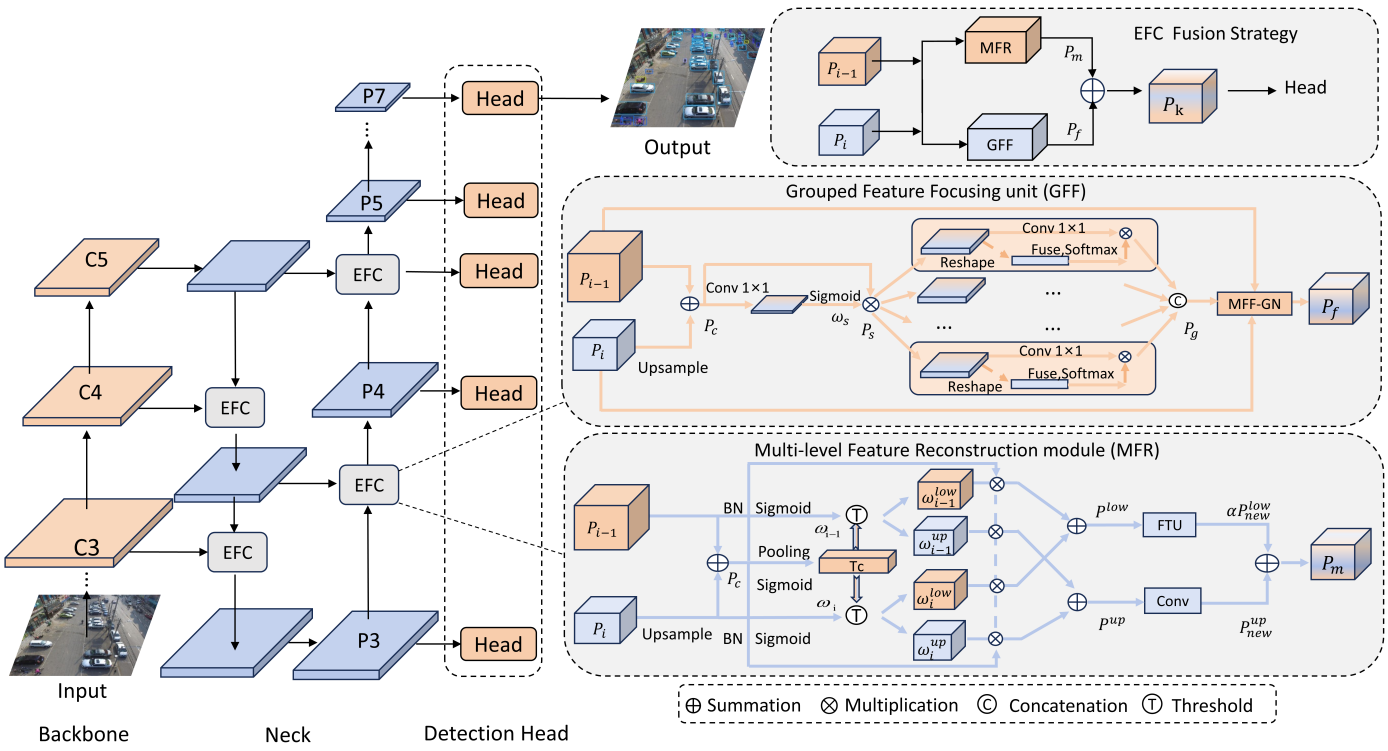

Fig. 2. The detailed structure of the FPN framework based on the enhanced inter-layer feature correlation lightweight fusion strategy (EFC). Given GFL as the base detector. It is noteworthy that this method can be flexibly applied to any detector based on the FPN architecture. The EFC replaces traditional fusion strategies and the large kernel convolutions in the neck with the Grouped Feature Focusing unit (GFF) and the Multi-level Feature Reconstruction module (MFR).

difficult localization of small objects in the network. Notably, our method can be directly applied to the aforementioned approaches.

C. Feature Enhancement and Fusion Methods

The construction of feature pyramids is a crucial step in many computer vision tasks and is an integral component of modern detectors, forming the foundation for addressing multi-scale problems. For smaller targets, the feature map often contains effective information from only a few or even just one pixel. Therefore, research on feature fusion methods is particularly important for accurately representing the feature information of small targets. The Feature Pyramid Network (FPN) builds a top-down pathway that combines features at various levels to achieve multi-scale feature fusion. PANet [24], based on FPN, introduces a bottom-up path, facilitating the fusion of high-resolution information with stronger semantic features. Subsequently, NAS-FPN [25] and BiFPN [26] are proposed to enhance the fusion of multi-scale features. Unlike many works focusing on cross-scale connections, A2-FPN [27] explores node operations on aggregated features, utilizing attention mechanisms to guide feature fusion. Although this method enhances detection performance, it significantly increases computational complexity. In this paper, we investigate the fundamental issues of feature fusion and achieve superior detection performance with a computational complexity far lower than that of the original FPN.

D. Lightweight Structural Design

Convolutional neural networks (CNNs) have achieved remarkable results in computer vision tasks, but this often comes with significant computational resource demands. A key goal is to attain better performance with limited computational resources. Techniques such as knowledge distillation and network pruning are commonly employed to lighten models. Additionally, there are efforts focused on lightweight backbone networks, such as MobileNet [28] and ShuffleNet [29], as well as lightweight detection heads [22], [23]. In contrast, our work focuses on the feature pyramid portion of the network. We address existing deficiencies in feature fusion by optimizing the feature fusion structure to efficiently utilize computational resources. This approach minimizes the fusion and extraction of ineffective features, achieving a balance between efficiency and complexity.

III. Method

In this section, we provide a detailed introduction to our proposed lightweight fusion strategy with enhanced inter-layer feature correlation (EFC), designed to optimize the fusion of features across different layers. The EFC consists of two main components: the Grouped Feature Focusing unit (GFF) and the Multi-level Feature Reconstruction module (MFR). The GFF enhances the correlation between adjacent features and focuses on key information. The MFR separates strong and weak spatial information, utilizing a lightweight convolutional

module to achieve precise feature transformation. This method reduces the extraction of irrelevant information while preserving crucial details of small objects in deep networks.

A. Grouped Feature Focus Unit

Spatial Concentration. To effectively combine the semantics of adjacent layers from the backbone network, which possess varying degrees of abstract semantic information, and to extract relevant feature information from different channels, we introduce a Grouped Feature Focus Unit (GFF). This unit enhances the correlation between features and improves the expression of information. As illustrated in Fig. 2, $P_i \in \mathbb{R}^{C_1 \times \frac{H}{2} \times \frac{W}{2}}$ and $P_{i-1} \in \mathbb{R}^{C \times H \times W}$ represent single-stage features from different stages. The low-resolution feature P_i is first upsampled using linear interpolation, followed by a 1×1 convolution to ensure that the channel number of the feature map remains consistent. This processed feature is then element-wise added with the high-resolution feature P_{i-1} to obtain the coarse feature $P_{\mathrm{c}} \in \mathbb{R}^{C \times H \times W}$. To refine this feature and acquire context-aware information, we use a 1×1 convolution to compress the feature into a single channel to aggregate spatial information, followed by sigmoid activation to generate spatial aggregation weights $\omega_{\mathrm{s}} \in \mathbb{R}^{1 \times H \times W}$. The feature $P_{\mathrm{s}} \in \mathbb{R}^{C \times H \times W}$ containing spatial information can be calculated as:

$$\omega_{\mathrm{s}} = \mathrm{Sigmoid}(\mathrm{Conv}(P_i \oplus P_{i-1})), \tag{1}$$

$$P_{\mathrm{s}} = (P_{\mathrm{c}} \otimes \omega_{\mathrm{s}}), \tag{2}$$

where $\otimes$ is element-wise multiplication, $\oplus$ is Element-wise summation. Sigmoid represents an activation function. Conv represents 1×1 convolutional layer.

Feature Correlation. To enhance the correlation between adjacent features, we divide the spatially aggregated feature P_{s} into n groups along the channel dimension, and perform feature interaction on a per-group basis. Specifically, we refine the feature information of adjacent channels within each group $[P_{\mathrm{s}}]^g_{n=i} \in \mathbb{R}^{\frac{C}{n} \times H \times W}$ using a convolution module. The global features from different channels in group $[P_{\mathrm{s}}]^g_{n=i}$ undergo the transformation to generate an attention mask ω_{g} that captures inter channel feature correlations. This mask ω_{g} is then applied to the refined feature. Finally, the features from each group are concatenated to form the aggregated and highly correlated adjacent features $P_g \in \mathbb{R}^{C \times H \times W}$. The entire computation process is as follows:

$$P_{gi} = (\mathrm{Softmax}(\mathcal{F}([P_{\mathrm{s}}]^g_{n=i}))) \otimes (\mathcal{N}([P_{\mathrm{s}}]^g_{n=i})), \tag{3}$$

$$P_g = P_{g1} \cup P_{g2} \cdots \cup P_{gi}, \tag{4}$$

where $\mathcal{F}$ and $\mathcal{N}$ represent the fused interaction layer and convolutional transformation Layer, respectively. $\cup$ is concatenation. P_{gi} represents each highly correlated feature group. Softmax is used as an activation function to generate the attention masks.

Spatial Mapping Normalization. Finally, we embed the grouped aggregated feature P_g into a normalization layer with multi-layered original feature fusion (MFF). We normalize feature P_g using its mean and standard deviation, thereby incorporating more spatial positional information from smaller targets. Through MFF-GN, we obtain feature P_{f} with strong feature correlation and enriched spatial information, which can be formulated as :

$$P_{\mathrm{f}} = \frac{P_g - \mathrm{mean}(P_i \oplus P_{i-1})}{\mathrm{std}(P_i \oplus P_{i-1})}, \tag{5}$$

where mean$(\cdot)$ and std$(\cdot)$ represent the mean and standard deviation, respectively.

By employing this approach, we fully leverage semantic information from adjacent layers and extract relevant features across different channels, thereby enhancing the overall feature representation.

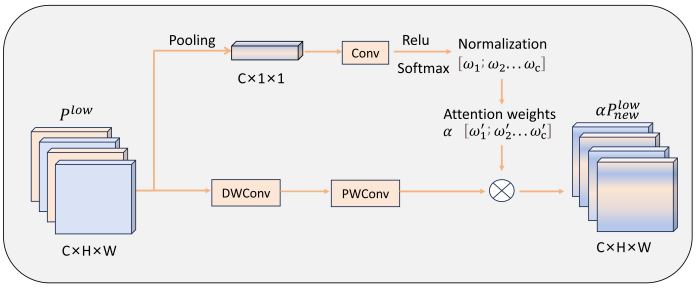

Fig. 3. An illustration of the structure of Feature Transformation Unit (FTU).

B. Multi-level Feature Reconstruction Module

To reduce the fusion and extraction of irrelevant features and minimize the loss of target information in deep networks, we design a Multi-level Feature Reconstruction Module (MFR). The purpose of reconstructing features is to separate the rich information from the weaker information contained in the feature maps from different stages of the backbone network and to process them independently. This allows the retention of abundant features to the greatest extent while requiring minimal computational resources to transform weaker features. Since information about small targets is particularly prone to loss during feature extraction and fusion, this process of reconstruction and independent transformation helps mitigate the loss of small target information to some extent.

Feature Separation. Specifically, as mentioned earlier, we obtain feature $P_{\mathrm{c}} \in \mathbb{R}^{C \times H \times W}$ from single-level features $P_i \in \mathbb{R}^{C_1 \times \frac{H}{2} \times \frac{W}{2}}$ and $P_{i-1} \in \mathbb{R}^{C \times H \times W}$ from different stages through operations such as upsampling, convolution, and element-wise addition. Next, we apply average pooling and the sigmoid function to generate information weights across each channel, which serve as the feature weight thresholds $T_{\mathrm{c}} \in \mathbb{R}^{C \times 1 \times 1}$. T_c can be formulated as :

$$T_{\mathrm{c}} = \mathrm{Sigmoid}(\mathrm{avg}(P_{\mathrm{c}})), \tag{6}$$

where avg$(\cdot)$ represent the average pooling.

The single-level features P_i and P_{i-1} are individually processed through Batch Normalization (BN) [35] and activated by the sigmoid function, generating unique weight information $\omega_i \in \mathbb{R}^{C \times H \times W}$ and $\omega_{i-1} \in \mathbb{R}^{C \times H \times W}$ at each spatial

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ART APPROACHES ON VISDRONE VALIDATION SET. '-' INDICATES THAT THE RESULT IS NOT REPORTED OR NOT PUBLICLY AVAILABLE.

Base Detector	Method	Backbone	mAP	AP_{50}	AP_{75}	FLOPs	Params
YOLOv3-SPP3 [30]	Baseline	Darknet-53	26.4%	-	-	284.10G	63.9M
ClusDet [31]	Baseline	ResNeXt-101	28.4%	56.2%	31.6%	-	-
Improved CascadeNet [32]	Baseline	ResNet-50	28.8%	47.1%	29.3%	-	-
DMNet [33]	Baseline	ResNeXt-101	29.4%	49.3%	30.6%	-	-
YOLC [34]	Baseline	ResNeXt-101	29.7%	52.4%	29.2%	-	-
Faster R-CNN [12]	Baseline	ResNet18	24.8%	43.6%	25.0%	322.25G	29.97M
	EFC(Ours)	ResNet18	**25.6%**	**44.8%**	**25.7%**	**301.36G**	**28.53M**
RetinaNet [15]	Baseline	ResNet50	20.2%	36.9%	19.5%	586.77G	42.74M
	QueryDet [22]	ResNet50	19.6%	35.7%	19.0%	-	-
	QueryDet-CSQ [22]	ResNet50	19.3%	35.0%	18.9%	-	-
	CEASC [23]	ResNet50	20.8%	35.0%	**27.7%**	**201.96G**	-
	EFC(Ours)	ResNet50	**23.3%**	**39.2%**	23.9%	482.25G	**40.22M**
GFL [16]	Baseline	ResNet18	28.4%	50.0%	27.8%	524.95G	42.52M
	MobileNet V2 [28]	MobileNet V2	28.5%	50.2%	28.1%	491.47G	-
	ShuffleNet V2 [29]	ShuffleNet V2	26.2%	46.6%	25.7%	488.94G	-
	CEASC [23]	ResNet18	28.7%	50.7%	28.4%	**150.18G**	-
	EFC(Ours)	ResNet18	**30.1%**	**52.1%**	**29.8%**	477.61G	**39.98M**

location, which indicates the importance of different feature maps.

$$\omega_i = \text{Sigmoid}(\text{BN}(P_i)), \tag{7}$$

$$\omega_{i-1} = \text{Sigmoid}(\text{BN}(P_{i-1})), \tag{8}$$

where BN stands for Batch Normalization [35].

Next, the weight informations ω_i and ω_{i-1} from different stages are then compared with the feature weight thresholds T_c to obtain attention maps that capture the strength of spatial information. Subsequently, strong and weak features from different layers are aggregated separately to yield the enriched features and weak features.

$$(\omega_i^{up}, \omega_i^{low}) = \text{Threshold}(\omega_i, T_\text{c}), \tag{9}$$

$$(\omega_{i-1}^{up}, \omega_{i-1}^{low}) = \text{Threshold}(\omega_{i-1}, T_\text{c}), \tag{10}$$

where we use a threshold function to separate strong and weak feature information.

Directional Fusion. The strong attention maps ω_i^{up} and ω_{i-1}^{up} are individually mapped onto feature P_c to and then fused to generate the enriched features. Similarly, the weak attention map is mapped onto P_c to generate the weak features. The entire computation process is as follows:

$$P^{up} = (\omega_i^{up} \otimes P_\text{c}) + (\omega_{i-1}^{up} \otimes P_\text{c}), \tag{11}$$

$$P^{low} = (\omega_i^{low} \otimes P_\text{c}) + (\omega_{i-1}^{low} \otimes P_\text{c}), \tag{12}$$

where $P^{up} \in \mathbb{R}^{C\times H\times W}$ represents the enriched features generated through reconstruction, and $P^{low} \in \mathbb{R}^{C\times H\times W}$ represents the weak features generated through reconstruction.

Feature Transformation. We transform the features P^{up} and P^{low} separately. For the enriched features, we apply a 1×1 convolution to generate feature maps P_{new}^{up} that display more detailed information. For the weak features P^{low} is fed into a feature transformation unit (FTU) designed to produce feature maps with richer semantic information using fewer computational resources. As Fig. 3 shows, we employ depthwise separable convolutions, which have lower computational and parameter overhead. Since depthwise separable convolutions disrupt inter-channel information flow, we generate feature modulation between channels. After the depthwise separable convolution operation, we perform weighted mapping to enhance information flow between channels. The weighted features α are processed through adaptive average pooling and convolution layers, which can be formulated as:

$$\alpha = \text{Softmax}(\mathcal{T}(\mathcal{A}(P^{low}))), \tag{13}$$

where $\mathcal{T}$ represents the convolutional transformation layer, and $\mathcal{A}$ represents the adaptive average pooling layer.

Level-wise Fusion. Finally, we merge the feature P_{new}^{low}, processed through the feature transformation unit, with the feature map P_{new}^{up}, which displays more detailed information, to generate the feature P_m. This feature contains both detailed information and cross-channel information exchange. The computation of P_m is as follows:

$$P_\text{m} = \alpha P_{new}^{low} + P_{new}^{up}. \tag{14}$$

Overall, we merge features from two different layers using a multi-level feature reconstruction module (MFR), resulting in enriched features with more detail while reducing computational resource usage. This approach enables the specific transformation of individual features, thereby minimizing the generation of redundant features.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON COCO TEST-DEV. '-' INDICATES THAT THE RESULT IS NOT REPORTED OR NOT PUBLICLY AVAILABLE.

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FLOPs	Params
Faster R-CNN [12]	ResNet-50	36.2%	59.1%	39.0%	18.2%	39.0%	48.2%	207.07G	41.53M
Mask R-CNN [11]	ResNet-50	38.2%	**60.3%**	41.7%	20.1%	41.1%	50.2%	260.14G	44.17M
RefineDet512 [36]	ResNet-101	36.4%	57.5%	39.5%	16.6%	39.9%	**51.4%**	-	-
FCOS [37]	ResNet-50	36.6%	56.0%	38.8%	21.0%	40.0%	47.1%	200.55G	32.02M
FSAF [38]	ResNet-50	37.4%	56.8%	39.8%	20.4%	41.1%	48.8%	282.35G	55.19M
RetinaNet [15]	ResNet-50	36.5%	55.4%	39.1%	20.4%	40.3%	48.1%	239.32G	37.74M
RetinaNet + EFC	ResNet-50	37.6%	56.8%	39.8%	22.1%	40.8%	48.5%	231.72G	32.79M
GFL [16]	ResNet-50	39.2%	56.9%	42.7%	21.9%	43.4%	51.2%	208.39G	32.22M
GFL+ EFC	ResNet-50	**40.2%**	58.1%	**43.8%**	**23.3%**	**44.0%**	**51.4%**	**202.90G**	**31.60M**

C. EFC as a Feature Fusion Strategy

Following the output of the Grouped Feature Focus Unit, feature P_{f} exhibits correlated features across different levels, focusing on perceptual spatial context. Feature P_{m} , generated by the Multi-level Feature Reconstruction Module, retains substantial information on small-scale targets and enhances semantic expression. The generated feature P_{k} results from integrating P_{f} and P_{m} at higher levels, ensuring a coherent representation of spatial and semantic information relevant to small targets. EFC replaces traditional, straightforward fusion operations such as simple Contact or Add methods for adjacent features.

D. Analysis on Complexities

Our EFC reduces the use of large kernel convolutions during the feature fusion stage, which helps to minimize the generation of redundant features. We analyze the theoretical parameter consumption, and the parameters of standard 3×3 convolution are calculated as:

$$P_{\mathrm{st}} = 3 \times 3 \times C_1 \times C_2, \tag{15}$$

where C_1 and C_2 represent the number of input and output feature channels, respectively.

EFC parameter consumption mainly occurs during the feature transformation and feature correlation stages. The parameter consumption of GFF and MFR, denoted as P_G and P_M respectively, is calculated as follows:

$$P_{\mathrm{G}} = 1 \times 1 \times C_1 \times 1 + n \times (1 \times 1 \times \frac{C_1}{n} \times \frac{C_2}{n}), \tag{16}$$

$$P_{\mathrm{M}} = 3 \times 3 \times C_1 \times 1 + 3 \times 1 \times 1 \times C_1 \times C_2, \tag{17}$$

where n represents the number of groups. In the experiment, we set the number of channels to $C_1 = C_2$, $n = 4$. The memory usage of our method is significantly smaller than that of standard convolution.

IV. EXPERIMENTS

In this section, we validate the effectiveness of our proposed method using three widely adopted benchmarks: VisDrone [7], UAVDT [8] and MS COCO [9]. We also conduct comprehensive ablation studies to thoroughly evaluate our contributions.

TABLE III
COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON UAVDT.'-' INDICATES THAT THE RESULT IS NOT REPORTED OR NOT PUBLICLY AVAILABLE.

Model	Backbone	AP	AP_{50}	AP_{75}
ClusDet [31]	ResNet-50	13.7%	26.5%	12.5%
GLSAN [39]	ResNet-50	17.0%	28.1%	18.8%
DREN [40]	ResNet-50	15.1%	-	-
GFL [16]	ResNet-18	16.9%	29.5%	17.9%
CEASC [23]	ResNet-18	17.1%	30.9%	17.8%
GFL+ EFC	ResNet-18	**18.0%**	**31.5%**	**18.9%**

A. Datasets and Metrics

VisDrone is a dataset characterized by a large number of small objects. It consists of 10,209 high-resolution (2,000 $\times$ 1,500) aerial images belonging to 10 categories (6,471 for training, 548 for validation, and 3,190 for testing). Since the evaluation server is currently closed, we are unable to test our method on the test dataset. Following previous work, we use 6,471 images for training and 548 images for testing. COCO is the most widely used dataset for general object detection, containing approximately 118K images in the train2017 set and 5K images in the val2017 set for training and validation. The UAVDT (Unmanned Aerial Vehicle Benchmark Object Detection and Tracking) dataset comprises 23,258 training images and 15,069 testing images. The image resolution is approximately 1,080 $\times$ 540 pixels. The dataset has been manually annotated for three vehicle categories: car, truck, and bus.

We utilize mean Average Precision (mAP) as the metrics for evaluating accuracy. We also use GFLOPs and Params to verify the complexity of the model.

B. Implementation Details

We implement our method based on the MMDetection framework [41]. All experiments are conducted on an NVIDIA GeForce RTX 3080 GPU. On VisDrone, we train all detectors with a batch size of 4 for 15 epochs, starting with an initial

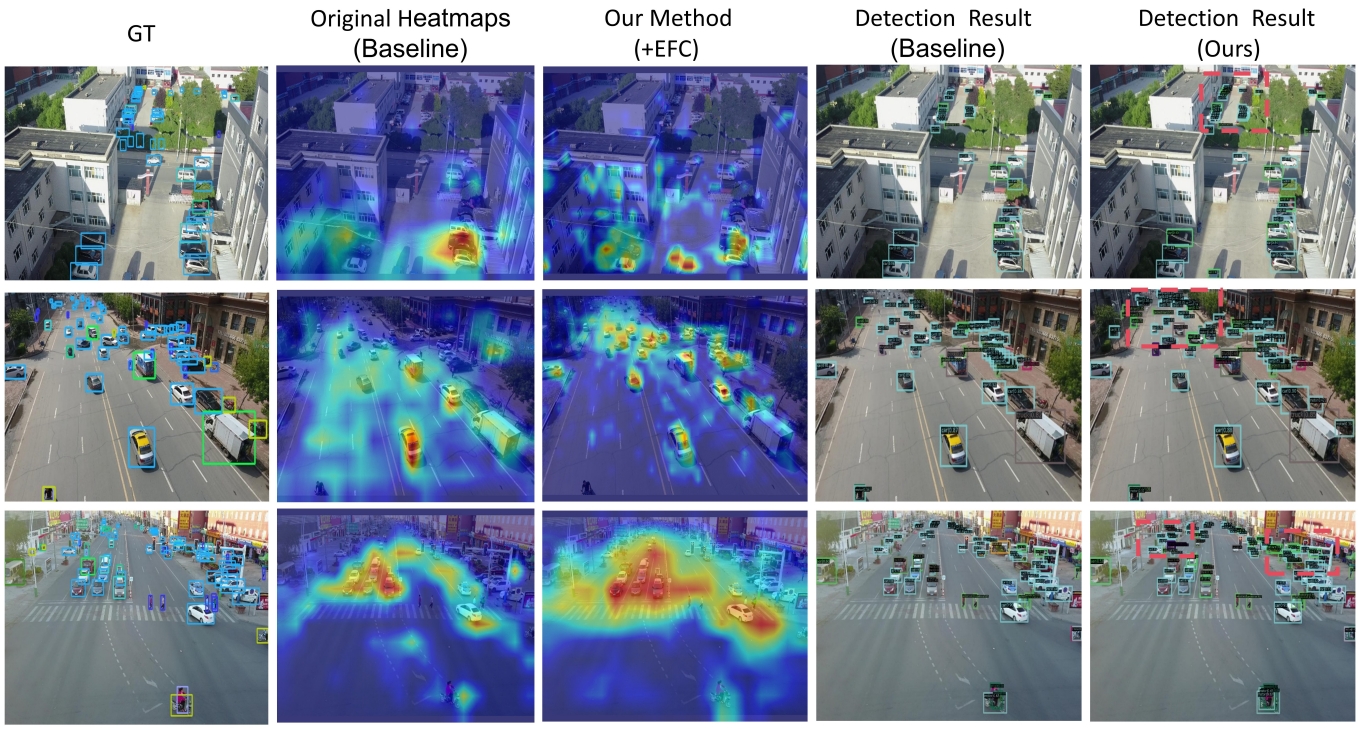

Fig. 4. Visualization of the detection results and heatmaps on VisDrone. The highlighted areas represent the regions that the network is focusing on. This demonstrates the superior performance of the proposed EFC method in small object detection.

learning rate of 0.01. This learning rate follows a linear warm-up strategy and is decreased by a factor of 0.1 at epochs 11 and 14. On COCO, we train detectors with a batchsize of 4 for 12 epochs with an initial learning rate of 0.01, and decrease it by 0.1 after 8 and 11 epochs. On UAVDT, we train models for 6 epochs using an initial learning rate of 0.01. After the 4 and 5 epochs, the learning rate is decreased by a factor of 10. The input image sizes are set to 1,333 $\times$ 800 on VisDrone and COCO. On UAVDT, the input image sizes are set to 1,024 $\times$ 540. We utilize GFL and RetinaNet as the base detectors. The output channels of the neck are set to 512 by default on the VisDrone dataset. All other parameters are configured according to the CEASC [23] guidelines.

C. Results on VisDrone Dataset

1) Evaluation on Different Detectors: EFC can be flexibly integrated into various state-of-the-art baseline detectors that utilize the FPN network. To validate the effectiveness of the method, we combine it with different state-of-the-art baseline detectors, including GFL and RetinaNet, using various backbone networks. The results, presented in Table I, demonstrate that by incorporating our method, both the number of parameters and the GFLOPs are reduced while detection accuracy improves compared to the baseline models. Specifically, by replacing the FPN in RetinaNet with our approach, the detection accuracy (AP) increased by 3.1%, while the computational load of the model decreased by 17.7%. Using GFL as the base detector, our method achieves 30.1 mAP, This underscores the superiority of our method in small object detection. It's worth noting that using our method reduces the GFLOPs by 42.7% and Params by 20.1% compared to the original FPN network's neck.

2) Comparison with State-of-the-art Methods: We report the results of comparing our proposed method with state-of-the-art lightweight methods, including lightweight backbone networks ShuffleNet V2 [29] and MobileNet V2 [28], as well as lightweight detection heads QueryDet [22] and CEASC [23]. For a fair comparison, our method employs the same data augmentation techniques as QueryDet and CEASC. As summarized in Table I, our method significantly improves detection accuracy while reducing Parameters and GFLOPs. Using GFL as the baseline, by replacing the lightweight backbone networks MobileNet V2 and ShuffleNet V2, their accuracy improvement is minimal. Using the lightweight detection head CEASC, although GFLOPs are greatly reduced, the detection accuracy improvement is minimal. Using GFL as the base detector, our method improves the detection accuracy mAP by 1.7% and 1.4% compared to the baseline model and CEASC, respectively. Our method greatly improves the detection accuracy of small targets through the fusion method on the neck end. This achieves a balance between model complexity and detection accuracy, and our method can be used simultaneously with other lightweight methods.

3) Visualization of Detection Results: To illustrate the advantages of our method more intuitively, we visualize the heatmaps of both the baseline model and our method in Fig. 4. From the results, it is evident that our method improves the receptive field for small objects, particularly in areas where the targets are dense and farther away from the camera, resulting in better detection results. This superior performance is due to our feature fusion process, which reduces the loss of

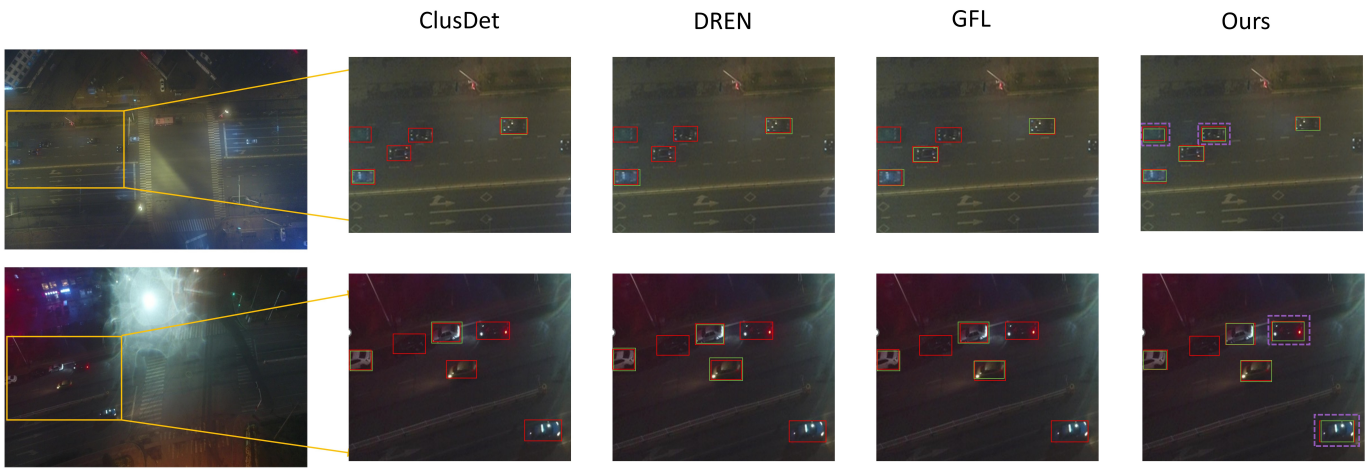

Fig. 5. Visualizations of the detection results of ClusDet, DREN, GFL and our proposed method under low light and similar background conditions on UAVDT. On the left side are the original images to be detected, and on the right side are the localized effect images detected by different methods. Red box is the ground truth, green box is the prediction, purple box represents the advantage detection results of our method.

small object information and increases the correlation between features.

D. Results on UAVDT Dataset

1) Comparison with State-of-the-art Methods: The UAVDT dataset contains numerous small targets and includes many low-light, complex background images, which can better reflect the performance of networks in small target detection. In Table III, we report the performance of our method on the UAVDT dataset. Using GFL as the baseline model, we achieve a 2.0% improvement in AP_{50}, surpassing many state-of-the-art methods. Compared to the latest lightweight method CEASC, our detection accuracy improves by 0.9% in AP, 0.6% in AP_{50}, and 1.1% in AP_{75}. The results demonstrate that our method performs well in object detection for UAV images.

2) Visualization Performance: Small target information is easily overwhelmed by environmental noise, resulting in extremely limited small target details. Moreover, under low-light conditions, targets tend to blend into the background, making detection challenging. Our approach focuses on integrating this limited small target information to achieve high-level feature expression. As illustrated in Fig. 5, we visualize the detection performance under low-light conditions and compare it visually with other state-of-the-art methods. From the figures, it is evident that ClusDet [31], DREN [40] and GFL struggle to detect some targets that resemble the background, thus being susceptible to background interference. In contrast, our method effectively utilizes the limited small target information to detect targets, demonstrating its superiority.

E. Results on COCO Dataset

1) Quantitative Evaluation: The COCO dataset is widely used and includes many small objects. We perform quantitative comparisons using our method across different baseline models and compare them with some state-of-the-art methods. The results on the COCO 2017 test-dev are summarized in Table II. Our method also shows a significant improvement in detection accuracy on this general dataset. It is worth noting that we use different baseline models to demonstrate that our method can be widely applied to various detectors and can further enhance the performance of state-of-the-art methods. Using RetinaNet and GFL as baseline models, our method increases AP_S by 1.7% and 1.4%, respectively. The experimental results indicate that our method is effective not only for small object detection in UAV images but also for general small object detection tasks.

TABLE IV
ABLATION ON GFF AND MFR WITH GFL AS THE BASE DETECTOR ON VISDRONE.

GFF	MFR	AP	AP_{50}	AP_{75}	Params	FLOPs
FPN		28.4%	50.0%	27.8%	42.52M	504.88G
PAFPN		28.9%	50.9%	28.6%	51.96M	523.49G
✓		29.6%	51.2%	29.3%	48.37M	523.32G
	✓	29.4%	51.0%	29.2%	**39.07M**	**475.95G**
✓	✓	**30.1%**	**52.1%**	**29.8%**	39.98M	477.61G

2) Visualization of Feature Map: To study the feature representation of our method, we visualize the first layer feature maps of the feature pyramid, comparing it with the baseline model. As shown in Fig. 6, our method integrates richer and more representative features compared to the baseline model.

F. Ablation Study

In this section, we conduct comprehensive ablation experiments to analyze the impact of the key components within the Grouped Feature Focusing unit (GFF) and the Multi-level Feature Reconstruction module (MFR). Notably, all ablation experiments are carried out on the VisDrone dataset, with GFL as the baseline model and ResNet18 as the backbone network.

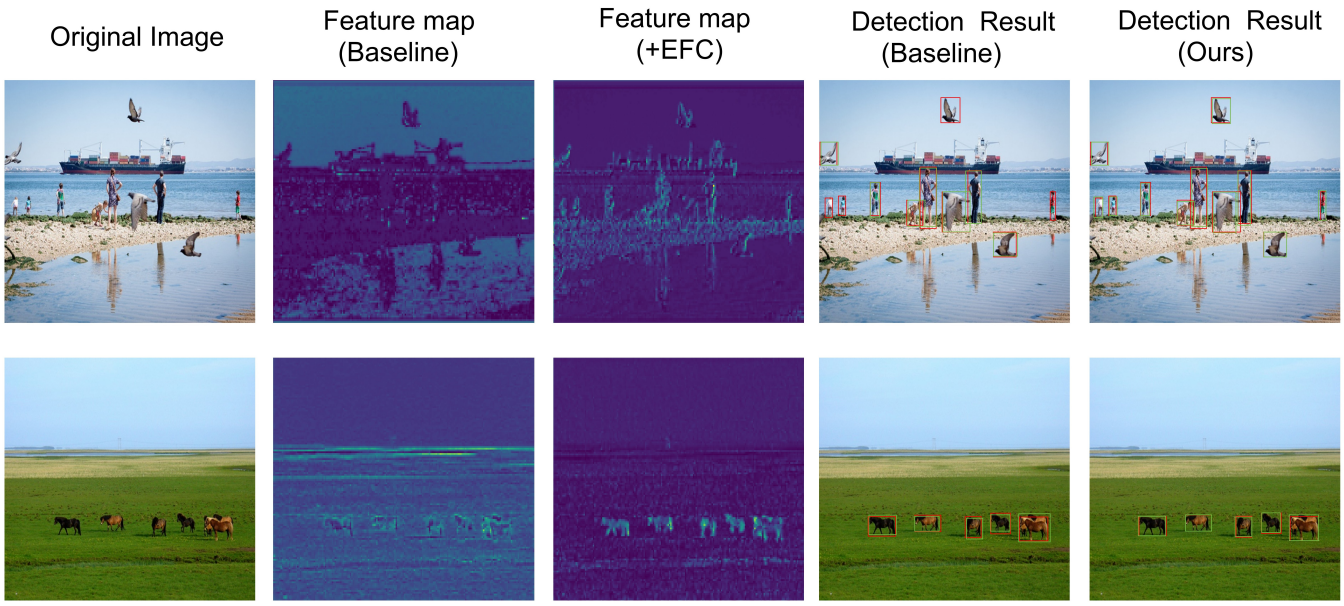

Fig. 6. Qualitative results comparison between RetinaNet and our method on COCO dataset. Red box is the ground truth, green box is the prediction. The feature maps are derived from the first stage of the feature pyramid. The brighter color represents that the model pays more attention to that area.

TABLE V
COMPARING DIFFERENT METHODS OF OBTAINING SPATIAL INFORMATION ON VISDRONE.

Method	AP	AP_{50}	AP_{75}	Params	FLOPs
Point-wise	29.6%	51.3%	29.1%	41.05M	484.35G
CBAM [42]	30.0%	51.9%	29.6%	40.28M	480.23G
3×3 Convolution	**30.2%**	51.8%	**29.8%**	49.44M	536.60G
Ours	30.1%	**52.1%**	**29.8%**	**39.98M**	**477.61G**

1) Effectiveness of Different Components: To validate the effectiveness of the EFC fusion strategy, we replace the traditional fusion structure with our proposed method and conduct ablation experiments on its two key components: the Grouped Feature Focusing unit (GFF) and the Multi-level Feature Reconstruction module (MFR). The baseline model employs the classic FPN and PAFPN structures. As illustrated in Table IV, each component of our method significantly enhances detection accuracy compared to the baseline model. The Grouped Feature Focusing unit enhances the correlation between features and strengthens the dependence of small objects across features, allowing for the perception of more semantic information. This leads to a 1.2% increase in AP compared to the baseline model. The Multi-level Feature Reconstruction module, which replaces the 3×3 convolutions in the neck, greatly reduces the computational resources consumed by the neck. Compared to PAFPN, the parameter count and GFLOPs decrease by 23.1% and 8.8%, respectively. Additionally, by reconstructing features, the loss of small object information in deep convolutions is mitigated, resulting in an improvement of 1.0% in detection accuracy.

2) Ablation Study on GFF: We evaluate the spatial attention mechanism for perceiving contextual information, the number of grouped units, and the MFF-GN structure within the Grouped Feature Focusing (GFF) unit separately. To capture contextual information from different layers, we use spatial attention for focusing. We compare this approach with several other context-aware methods, such as Point-wise, CBAM [42], and standard 3×3 convolutions. As shown in Table V, utilizing the Spatial Attention Module highlights its advantages in accuracy and resource utilization compared to the mentioned methods. The standard 3×3 convolution achieves almost the same detection accuracy, however, this comes at the cost of high computational resources. Our method achieves the best detection performance with the lowest GFLOPs and Params.

Next, we investigate the impact of varying the number of groups on detection accuracy, as shown in Table VI. The results indicate that setting the number of groups to 4 achieves a favorable balance between detection accuracy and efficiency.

To enhance spatial mappings from features across different layers and improve the representation of features for small objects in the original space, we employ MFF-GN for normalization. To evaluate the performance of MFF-GN, we compare it with other normalization techniques such as Batch Normalization (BN) [35], Group Normalization (GN) [43], Switchable Normalization (SN) [44], Instance Normalization (IN) [45], and standard 1×1 convolutions, as well as the baseline detector without normalization. As shown in Table VII, compared to other normalization methods, MFF-GN achieves the best detection accuracy. It outperforms GN by 0.2% in AP and also demonstrates advanced computational resource utilization.

3) Ablation Study on MFR: We transform the reconstructed strong and weak features separately, aiming to preserve the

TABLE VI
ABLATION EXPERIMENTS ON VISDRONE DATASET WITH VARYING NUMBERS OF GROUPS.

Number of Groups	AP	AP_{50}	AP_{75}	FLOPs	Params
g = 2	30.0%	51.9%	29.7%	479.25G	40.20M
g = 4	**30.1%**	**52.1%**	**29.8%**	477.61G	39.98M
g = 6	30.0%	52.0%	29.7%	477.21G	39.83M
g = 8	29.9%	51.8%	29.5%	**476.79G**	**39.71M**

TABLE VII
COMPARISON OF DIFFERENT NORMALIZATION METHODS ON VISDRONE.

Method	AP	AP_{50}	AP_{75}	FLOPs
Group Normalization [43]	29.9%	52.0%	29.6%	478.53G
Batch Normalization [35]	29.5%	51.2%	29.1%	**476.14G**
Instance Normalization [45]	29.8%	51.7%	29.5%	485.46G
Switchable Normalization [44]	29.8%	51.8%	29.6%	481.28G
MFF-GN (Ours)	**30.1%**	**52.1%**	**29.8%**	477.61G

TABLE VIII
COMPARISON OF DIFFERENT LIGHTWEIGHT CONVOLUTION METHODS FOR FEATURE TRANSFORMATION ON VISDRONE.

Method	AP	AP_{50}	AP_{75}	FLOPs	Params
DWConv [46]	29.5%	51.7%	29.4%	**470.60G**	**38.95M**
GConv [47]	29.8%	52.0%	29.6%	506.98G	43.65M
PConv [48]	29.7%	51.9%	29.6%	493.45G	41.86M
3×3 Standard Conv	**30.1%**	**52.2%**	29.7%	536.47G	48.37M
FTU (Ours)	**30.1%**	52.1%	**29.8%**	477.61G	39.98M

TABLE IX
COMPARING THE IMPACT OF DIFFERENT COMPONENT CONFIGURATIONS IN STRONG AND WEAK FEATURE TRANSFORMATION ON VISDRONE. P^{up} AND P^{low} REPRESENT THE STRONG AND WEAK FEATURES.

Method	AP	AP_{50}	AP_{75}	Params
P^{up}(FTU), P^{low}(FTU)	29.9%	51.8%	29.5%	41.08M
P^{up}(Conv), P^{low}(Conv)	29.6%	51.4%	29.1%	**38.93M**
P^{up}(FTU), P^{low}(Conv)	29.7%	51.5%	29.3%	39.98M
P^{low}(FTU), P^{up}(Conv)	**30.1%**	**52.1%**	**29.8%**	39.98M

rich features and convert the weaker ones. This approach helps retain more subtle information of small objects while saving computational resources. For strong features, we refine them using 1×1 convolutions, while weak features are transformed using lightweight computational methods. To validate the superiority of our transformation module, we compare FTU with other lightweight convolution modules, including depthwise separable convolutions [46], group convolutions [47], and partial convolutions [48]. As shown in Table VIII, the table unequivocally demonstrates that the FTU module surpasses other methods in terms of performance while consuming fewer computational resources, achieving a commendable balance between accuracy and efficiency. Specifically, compared to DWConv, GConv, and PConv, the AP is improved by 0.6%, 0.3%, and 0.4%, respectively. Moreover, the 3×3 standard convolution and FTU achieve similar levels of accuracy, yet with significantly increased GFLOPs and Params.

To highlight the advantages of separately transforming strong and weak features, we design a series of variant experiments to demonstrate that our configuration is optimal. As shown in Table IX, weak features are enhanced through FTU to improve feature information, while strong features are processed with 1×1 convolutions to reveal more detailed information. This process achieves high-level feature transformation, which helps enhance the representation of small object information.

V. CONCLUSION

In this work, we propose an enhanced inter-layer feature correlation lightweight fusion strategy (EFC), it includes the Grouped Feature Focusing unit (GFF) to increase the correlation between features and enhance the spatial mapping of small object information. We also introduce the Multi-level Feature Reconstruction module (MFR), which aims to separate and reconstruct features from different layers, utilizing lightweight operations for directed feature transformation. This approach reduces the loss of small object information in deep networks and minimizes the extraction of irrelevant features. Notably, our proposed method can be flexibly integrated into the FPN network. Extensive experimental results on VisDrone, UAVDT and COCO demonstrate the effectiveness of EFC in small object detection and significantly reduce the computational resources at the neck.

ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Chongqing, China (Grant No. cstc2021jcyj-msxmX1130).

REFERENCES

[1] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[2] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "Ffca-yolo for small object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[3] V. Gagliardi, F. Tosti, L. Bianchini Ciampoli, M. L. Battagliere, L. D'Amato, A. M. Alani, and A. Benedetto, "Satellite remote sensing and non-destructive testing methods for transport infrastructure monitoring: Advances, challenges and perspectives," *Remote Sensing*, vol. 15, no. 2, p. 418, 2023.

[4] B. Huang, J. Li, J. Chen, G. Wang, J. Zhao, and T. Xu, "Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 2852–2865, 2024.

[5] H. Zhao, J. Chen, L. Wang, and H. Lu, "Arkittrack: a new diverse dataset for tracking using mobile rgb-d data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5126–5135.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[7] P. Zhu, L. Wen, D. Du, X. Bian, H. Ling, Q. Hu, Q. Nie, H. Cheng, C. Liu, X. Liu *et al.*, "Visdrone-det2018: The vision meets drone object detection in image challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[8] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[10] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "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*, 2023, pp. 7464–7475.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[16] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.

[17] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] H. Zhang, S. Wen, Z. Wei, and Z. Chen, "High-resolution feature generator for small ship detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] J. Li, S. Dong, L. Ding, and T. Xu, "Mssvt++: Mixed-scale sparse voxel transformer with center voting for 3d object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[20] J. He, Y. Wang, L. Wang, H. Lu, B. Luo, J.-Y. He, J.-P. Lan, Y. Geng, and X. Xie, "Towards deeply unified depth-aware panoptic segmentation with bi-directional guidance learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4111–4121.

[21] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[22] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 13 668–13 677.

[23] B. Du, Y. Huang, J. Chen, and D. Huang, "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*, 2023, pp. 13 435–13 444.

[24] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[25] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7036–7045.

[26] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[27] M. Hu, Y. Li, L. Fang, and S. Wang, "A2-fpn: Attention aggregation based feature pyramid network for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 343–15 352.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[29] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[30] P. Zhang, Y. Zhong, and X. Li, "Slimyolov3: Narrower, faster and better for real-time uav applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[31] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8311–8320.

[32] X. Zhang, E. Izquierdo, and K. Chandramouli, "Dense and small object detection in uav vision based on cascade network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[33] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 190–191.

[34] C. Liu, G. Gao, Z. Huang, Z. Hu, Q. Liu, and Y. Wang, "Yolc: You only look clusters for tiny object detection in aerial images," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[36] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[37] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9627–9636.

[38] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 840–849.

[39] S. Deng, S. Li, K. Xie, W. Song, X. Liao, A. Hao, and H. Qin, "A global-local self-adaptive network for drone-view object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 1556–1569, 2020.

[40] J. Zhang, J. Huang, X. Chen, and D. Zhang, "How to fully exploit the abilities of aerial image detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[41] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[43] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[44] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li, "Differentiable learning-to-normalize via switchable normalization," *arXiv preprint arXiv:1806.10779*, 2018.

[45] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[46] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *arXiv preprint arXiv:1403.1687*, 2014.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[48] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: Chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 021–12 031.

Yao Xiao received the B.E. degree from China Jiliang University, Hangzhou, Zhejiang, China, in 2022, where he is currently pursuing the M.S. degree in optical engineering at the School of Optics and Photonics, Beijing Institute of Technology, China. His research interests include object detection and related computer vision problems.

Tingfa Xu received the Ph.D. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Changchun, China, in 2004. He is currently a Professor with the School of Optics and Photonics, Beijing Institute of Technology, Beijing, China. His research interests include optoelectronic imaging and detection and hyperspectral remote sensing image processing.

Xin Yu received the B.E. degree from ShanDong University, Qingdao, ShanDong, China, in 2022, where he is currently pursuing the M.S. degree in optical engineering at the School of Optics and Photonics, Beijing Institute of Technology, China. His research interests include deep learning and object detection.

Yuqiang Fang received the Ph.D. degree in control science and engineering from the National University of Defense Technology, Changsha, China, in 2015. He is currently an Associate Professor with Space Engineering University, Beijing, China. His research interests include machine learning, computer vision, and data mining.

Jianan Li is currently an assistant professor at School of Optics and Photonics, Beijing Institute of Technology, Beijing, China, where he received his B.S. and Ph.D. degree in 2013 and 2019, respectively. From July 2015 to July 2017, he worked as a joint training Ph.D. student at National University of Singapore. From October 2017 to April 2018, he worked as an intern at Adobe Research. His research interests mainly include computer vision and real-time image/video processing.