Towards Region-Adaptive Feature Disentanglement and Enhancement for Small Object Detection

Yanchao Bi[1] , **Yang Ning**[1,*] , **Xiushan Nie**[1,*] , **Xiankai Lu**[2] , **Yongshun Gong**[2] and **Leida Li**[3]
[1]School of Computer Science and Technology, Shandong Jianzhu University, Jinan, China
[2]School of Software, Shandong University, Jinan, China
[3]School of Artificial Intelligence, Xidian University, Xi'an, China
2022110101@stu.sdjzu.edu.cn, ningyang20@sdjzu.edu.cn, niexsh@hotmail.com,
carrierlxk@gmail.com,ysgong@sdu.edu.cn, ldli@xidian.edu.cn

Abstract

Current feature fusion strategies often fail to adequately account for the influence of activation intensity across different scales on small object features, which impedes the effective detection of small objects. To address this limitation, we propose the Region-Adaptive Feature Disentanglement and Enhancement (RAFDE) strategy, which improves both downsampling and feature fusion by leveraging activation intensity variations at multiple scales. First, we introduce the Boundary Transitional Region-enhanced Downsampling (BTRD) module, which enhances boundary transitional regions containing both strongly and weakly activated features, thereby mitigating the loss of crucial boundary information for small objects. Second, we present the Regional-Adaptive Feature Fusion (RAFF) module, which adaptively disentangles and fuses co-activated and uni-activated regions from adjacent levels into the current level, effectively reducing the risk of small objects being overwhelmed. Extensive experiments on several public datasets demonstrate that the RAFDE strategy is highly effective and outperforms state-of-the-art methods. The code is available at https://github.com/b-yanchao/RAFDE.git.

1 Introduction

Object detection using Unmanned Aerial Vehicles (UAVs) plays a critical role in various applications, such as remote sensing [He *et al.*, 2024] and autonomous driving [He *et al.*, 2023]. However, the prevalence of small objects in UAV imagery poses a substantial challenge for models in extracting effective features, leading to a significant performance disparity between small and regular objects [Khanam R, 2024].

To address this challenge, multi-scale feature fusion methods leverage the distinct receptive field properties at each level by combining feature maps from different scales [Lin *et al.*, 2017; Yang *et al.*, 2024]. However, due to the limited features of small objects, these methods often result in weak activations, which increases the likelihood of losing critical information with existing downsampling strategies [Hou *et al.*, 2020]. Furthermore, directly fusing multi-scale features from adjacent levels can cause strong activations from medium to large objects to overwhelm those of small objects at the current level [Huang *et al.*, 2024b]. This exacerbates the loss of small object features and further diminishes the availability of learnable features for small objects (Fig. 1(a)).

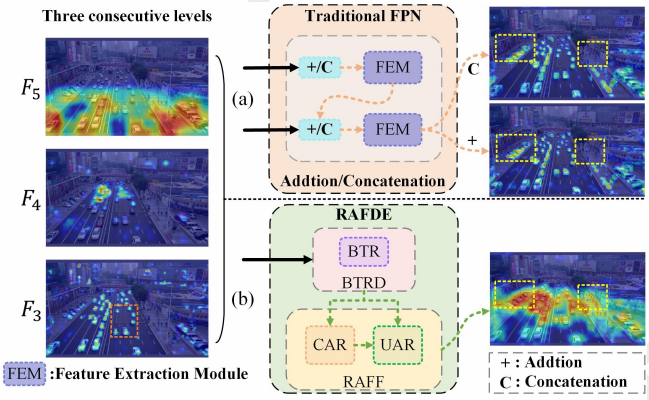

Figure 1: Our main motivation. (a) Traditional methods (*e.g.*, FPN) typically rely on "addition" or "concatenation" to integrate adjacent levels, which often fail to adequately account for the activation intensity differences across multiple scales. As a result, the fused feature maps exhibit poor activation for small objects. (b) To address this issue, we propose a new feature fusion strategy that consists of two operations: BTRD (BTR) and RAFF (CAR and UAR). The fused feature map has a stronger response for small objects.

To address the challenge that existing downsampling and feature fusion strategies lead to significant loss of small object details, thereby resulting in suboptimal performance for small objects, we have investigated the activation intensity of objects at different scales across adjacent levels. Based on these insights, we propose the Region-Adaptive Feature Disentanglement and Enhancement (RAFDE) strategy, which effectively reduces the loss of small object detail features. Specifically, through visualization of the feature maps generated by the backbone, we observed that small objects tend to activate only the central region due to their limited feature representation. This increases the risk of losing small objects during the backbone's downsampling process (see the orange box in F_3). To mitigate this, we introduce the Boundary Tran-

* Corresponding author: Yang Ning and Xiushan Nie.

sitional Region-enhanced Downsampling (BTRD) module. This module disentangles and enhances the boundary transitional regions, which contain both strongly and weakly activated features, effectively reducing the loss of critical boundary features for small objects during the downsampling process (see the BTRD in Fig. 1(b)).

Through the methods outlined above, we obtained multi-scale feature maps that effectively activate a larger number of small objects. However, there remains a need for a suitable strategy to fuse these features effectively. While extensive research on multi-scale feature fusion exists, most methods rely on addition or concatenation to combine feature maps, and still struggle with the loss of small object details (see yellow boxes in Fig. 1(a)). In multi-scale feature fusion, addition can enhance and filter critical features but may result in the loss of fine details. In contrast, concatenation preserves more details but may overwhelm small object features and introduce redundant information [Xiao *et al.*, 2024]. To address these challenges, we propose the Region-Adaptive Feature Fusion (RAFF) module, which disentangles co- and uni-activated regions across adjacent levels and applies addition and concatenation strategies for fusion, respectively. This approach effectively mitigates the risk of small object features being overwhelmed. Finally, we integrate this module to replace the Feature Pyramid Network in fusing feature maps from three adjacent levels, significantly reducing the loss of small object features after fusion (see yellow boxes in Fig. 1(b)).

The RAFDE is a simple yet effective method that can be seamlessly integrated with various convolution-based detectors. It significantly enhances model performance while maintaining a minimal increase in computational cost and substantially reducing the number of parameters. In summary, the key contributions of our work are reflected in three main aspects:

- We introduce the Boundary Transitional Region-Enhanced Downsampling (BTRD) module, which enhances boundary transitional regions that contain both strongly and weakly activated features, thereby reducing the loss of critical boundary details for small objects.
- We propose the Region-Adaptive Feature Fusion (RAFF) module, which disentangles co-activated and uni-activated regions across adjacent levels, selectively enhancing and fusing key features to mitigate the risk of small objects being overwhelmed.
- We evaluated the RAFDE on the VisDrone and Drone-vs-Bird datasets, achieving mAP improvements of 3.4% and 2.3%, respectively. The results demonstrate that RAFDE significantly enhances small object detection across different scenarios, underscoring its effectiveness in UAV detection.

2 Relate Work

2.1 Unmanned Aerial Vehicle Object Detection

In computer vision, detecting objects in UAV images is a challenging yet crucial task, as these objects typically exhibit a long-tailed distribution and are predominantly small. Due to the limited and densely packed nature of object features, existing models often struggle to extract effective representations, resulting in suboptimal performance.

In recent years, deep learning methods have made significant progress in addressing the challenges associated with small object detection. For instance, data augmentation techniques, such as copy-pasting, scaling, and cropping, increase the diversity and quantity of small object samples in UAV scenes, thereby enhancing dataset size [Meethal *et al.*, 2023; Wang *et al.*, 2023]. However, while these methods are effective for specific datasets, they lack generalizability due to their heavy dependence on particular data sources. Contextual learning strategies [Du *et al.*, 2023; Sun *et al.*, 2022; Huang *et al.*, 2024b] improve object and scene classification and localization by leveraging environmental and inter-object relationships. Nonetheless, not all objects have clear contextual information, and the use of incorrect context can negatively impact performance.Additionally, alternative methods [Dai *et al.*, 2024; Lei *et al.*, 2023; Moser *et al.*, 2024], including improvements in loss functions and attention mechanisms, have also boosted small object detection performance.

The widely used multi-scale feature fusion module effectively combines high- and low-level features, thereby improving the representation of small objects. However, traditional fusion strategies may overwhelm small object features by failing to account for the impact of large object activations in high-level features on small object activations in low-level features. To overcome these limitations, we propose the RAFDE strategy, a plug-and-play solution designed to reduce the risk of small object features being overwhelmed and lost, thereby improving small object detection performance. Additionally, we introduce a boundary transitional region-enhanced downsampling technique for the backbone, which generates multi-scale feature maps that activate more small object features. This approach significantly reduces model parameters while improving overall performance.

2.2 Multi-scale Object Detection Strategy

Multi-scale detection strategies address object scale variations by integrating high-level features, which capture richer semantics and larger receptive fields, with low-level features that preserve finer details, thereby enhancing small object detection.

Traditional computer vision methods typically extract features at a single scale, limiting their ability to detect objects of varying sizes or handle scenes with different proportions. To overcome this limitation, researchers developed the Feature Pyramid Network (FPN) [Lin *et al.*, 2017; Li *et al.*, 2024], which combines feature maps from multiple scales. By merging these maps, FPN leverages both detailed and semantic information, improving small object detection and addressing the constraints of single-scale features. However, this approach lacks effective integration between high-level and low-level features. To address this, the PAFPN [Liu *et al.*, 2018b] extended FPN by introducing a bottom-up path, allowing high-level features to retrieve finer details from low-level ones. Despite this, a significant semantic gap remains between high-level and low-level features, complicating the selection of a fusion method that minimizes the loss of small object features. For instance, AFPN [Yang *et*

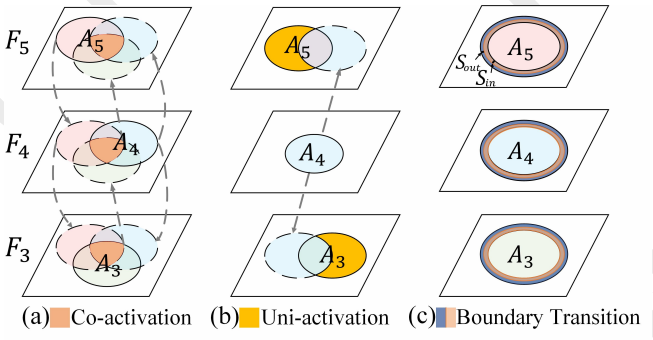

Figure 2: The definition of disentangled regions. (a) shows the co-activated regions with strongly activated features across three levels. (b) illustrates the uni-activated regions with strongly activated features in the adjacent levels and weakly activated features in the middle level. (c) depicts the boundary transition regions containing both strongly and weakly activated features. Note that the elliptical areas in the figure represent strongly activated features, while the other white areas indicate weakly activated features.

al., 2024] merges adjacent low-level features first, then gradually incorporates high-level features to reduce the semantic gap between non-adjacent levels. To further enhance multi-scale fusion, EFC [Xiao *et al.*, 2024] addresses the inadequacies of simple concatenation or addition by emphasizing the contextual relationships between features at each level, improving feature relevance.

While current methods effectively combine high-level semantics with low-level details, they often ignore that activations at different scales correspond to objects of varying sizes. This can lead to activations from larger objects overshadowing smaller ones during direct fusion, limiting model performance. To address these challenges, we propose a new feature fusion module that considers activation intensity across different scales, reducing the risk of small objects being overwhelmed during fusion process.

3 Method

3.1 Preliminary

To facilitate the understanding of our subsequent paper, we first provide the mathematical definition of the co-activated region, the uni-activated region, and the boundary transitional region, as shown in Fig. 2, respectively.

Consider three consecutive feature maps, F_3, F_4, and F_5, where each element $f_{i,j}^k$ represents the activation intensity at position (i, j) on the k-th feature map, with i and j as the row and column indices, respectively, and $k \in \{3, 4, 5\}$. Given a threshold T, if the activation intensity $f_{i,j}^k$ exceeds this threshold, the position (i, j) is considered strongly activated on the corresponding feature map [Selvaraju *et al.*, 2017].

Specifically, let A_k denote the set of all normalized strong activation features on the $k-th$ feature map. Then, the following definition holds:

$$A_k = \{(i, j) | f_{i,j}^k \geq T, (i, j) \in F_k\}, k \in \{3, 4, 5\}. \quad (1)$$

Thus, as shown in Fig. 2(a), the **co-activated region $\mathbf{\Phi}$** can be defined as:

$$\mathbf{\Phi} = \{(i, j) | f_{i,j}^3 \geq T \wedge f_{i,j}^4 \geq T \wedge f_{i,j}^5 \geq T\} = A_3 \cap A_4 \cap A_5. \quad (2)$$

Then, with the help of the weakly activated region of A_4 shown in Fig. 2(b), the **uni-activated region $\mathbf{\Omega}$** can be defined as:

$$\mathbf{\Omega} = (A_3 \cap (1 - A_4)) \cup (A_5 \cap (1 - A_4)). \quad (3)$$

Finally, as shown in Fig. 2(c), for the feature map F_5 that simultaneously contains strongly and weakly activated regions, we define the **boundary transitional region $\mathbf{\Psi}$** as:

$$\mathbf{\Psi} = A_k^{+s_{in}} \cap A_k^{+s_{out}}, k \in \{3, 4, 5\}, \quad (4)$$

where $A_k^{+s_{in}}$ and $A_k^{+s_{out}}$ denote the operations of expanding region A_k outward and inward by widths s_{in} and s_{out}, respectively. $s_{in} + s_{out}$ defines the width of the boundary transitional region, *i.e.*, the pooling kernel size.

3.2 Regional Adaptive Feature Fusion Module

In multi-scale feature fusion, it is important to acknowledge that feature maps at different levels have distinct receptive fields, which activate objects of varying scales [Lin *et al.*, 2017]. However, existing fusion strategies that simply merge adjacent levels often result in large objects in high-level features overwhelming small objects in low-level features.

To address these challenges, we propose the Region-Adaptive Feature Fusion (RAFF) module. The module carefully disentangles the co-activated regions and uni-activated regions across adjacent levels, effectively integrating the unique features from adjacent levels into the current level by leveraging the advantages of both fusion strategies. Specifically,as illustrated in the RAFF module in Fig. 3, we first align the high-level features F_{i+1}^o and low-level features F_{i-1}^o with the dimensions of the current-level feature map F_i. Specifically, the transformations of F_{i+1} and F_{i-1} are defined as follows:

$$F_{i+1} = UP(F_{i+1}^o),\ F_{i-1} = DWConv(F_{i-1}^o), \quad (5)$$

where $UP(\cdot)$ denotes the upsampling operation, and $DWConv(\cdot)$ refers to depthwise separable convolution, which reduces both computational cost and parameter count while maintaining performance.

We then apply the channel attention mechanism to emphasize relevant channel information, compressing it into a single-channel feature map. As a result, the activation maps for the three adjacent levels are represented as $\delta_{i-1}, \delta_i, \delta_{i+1}$:

$$\delta_{i+k} = F^{N \to 1}(ECA(F_{i+k})),\ k = -1, 0, 1, \quad (6)$$

where $ECA(\cdot)$ refers to the efficient channel attention mechanism [Wang *et al.*, 2020].

For the low-level features, we enhance δ_{i-1} and δ_i using the spatial attention mechanism [Woo *et al.*, 2018], resulting in their enhanced features η_{i-1} and η_i, respectively. Specifically, referring to Eq. (2), the **co-activated regions $\mathbf{\Phi_i^s}$** and **$\mathbf{\Phi_i^d}$** for the low- and high-level feature maps, respectively, can be approximated as:

$$\Phi_i^s = \sigma(\eta_{i-1} \times \eta_i),\ \Phi_i^d = \sigma(\delta_{i+1} \times \delta_i). \quad (7)$$

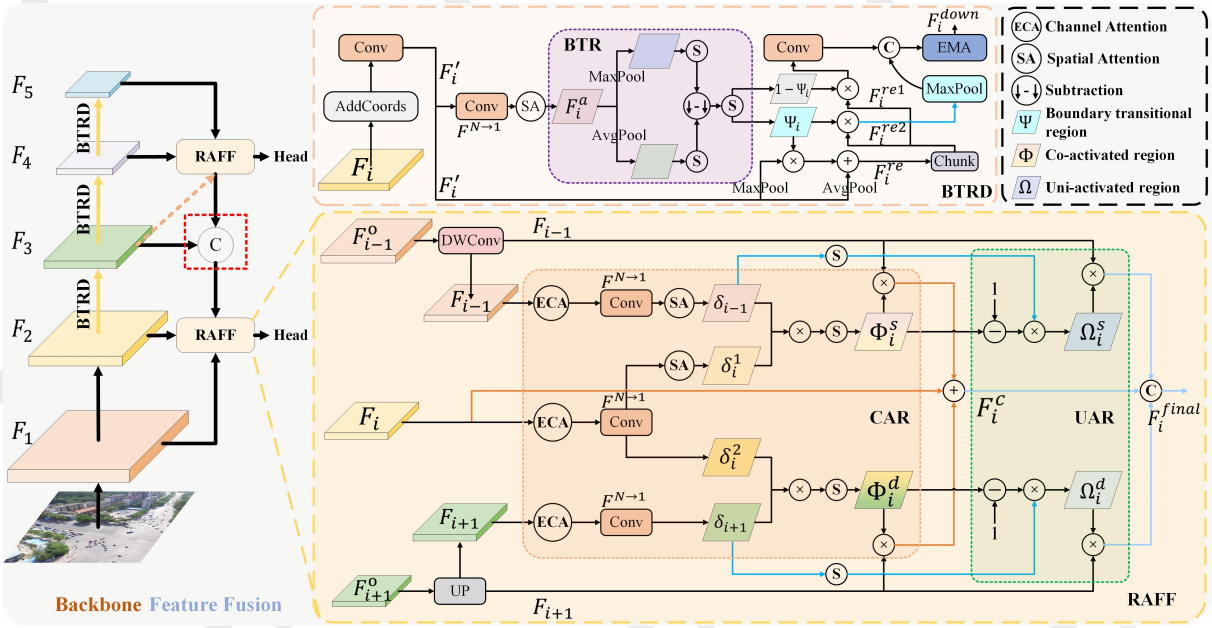

Figure 3: The architecture of the proposed RAFDE method consists of two key modules. The BTRD module focuses on enhancing object boundaries to minimize the loss of critical features for small objects, generating multi-scale feature maps that activate more small objects. Next, the RAFF module integrates distinctive features from different scales, effectively reducing the risk of small object features being overwhelmed. To preserve high-level semantic information, we use a concatenation operation to fuse these features into F_3 (red dotted box), significantly improving small object detection performance.

For the features within the co-activated regions, we apply addition to enhance them, reducing the risk of introducing redundant features and the potential for activated features to be overwhelmed. Consequently, the enhanced features F_i^c from the co-activated regions of the adjacent three levels can be defined as:

$$F_i^c = F_i + F_{i-1} \times \Phi_i^s + F_{i+1} \times \Phi_i^d. \tag{8}$$

Following the definition in Eq. (3), we can further obtain the approximated **uni-activated regions Ω_i^s** and **Ω_i^d** in the low- and high-level feature maps, respectively, as follows:

$$\Omega_i^s = \sigma(\delta_{i-1}) \times (1 - \Phi_i^s),\ \Omega_i^d = \sigma(\delta_{i+1}) \times (1 - \Phi_i^d). \tag{9}$$

For the features within the uni-activated regions, we apply concatenation to enhance them, thereby reducing the loss of unique features in adjacent levels. Therefore, the final feature map after region-adaptive fusion F_i^{final} is represented as:

$$F_i^{final} = F_i^c \oplus (F_{i-1} \times \Omega_i^s) \oplus (F_{i+1} \times \Omega_i^d). \tag{10}$$

The RAFF module effectively disentangles co-activated and uni-activated regions from adjacent levels, seamlessly fusing distinctive features from these levels into the current level through addition and concatenation. By substituting the traditional feature pyramid structure [Liu *et al.*, 2018b] with the RAFF module, we achieve better integration of critical information across multiple feature scales, while mitigating the risk of larger object features overwhelming the details of small objects.

3.3 Boundary Transitional Region-enhanced Downsampling

Due to their limited features, small objects often produce weak activations, which increases the likelihood of losing detailed information during downsampling in existing backbone networks, particularly at critical boundaries [Hou *et al.*, 2020; Chen *et al.*, 2023]. This makes it challenging to capture sufficient and effective features for small objects, ultimately resulting in poor detection performance.

To address the aforementioned challenges, we propose the Boundary Transitional Region-enhanced Downsampling (BTRD) module, which enhances the boundary transitional regions that contain both strongly and weakly activated features. This approach minimizes the loss of crucial information related to small objects during downsampling. Specifically, as shown in Fig. 3, we utilize coordinate convolution to integrate positional information into the feature map F_i, resulting in the enhanced feature map F_i':

$$F_i^{'} = Conv(AddCoords(F_i)), \tag{11}$$

where $AddCoords(\cdot)$ denotes the coordinate convolution operation [Liu *et al.*, 2018a] used to embed coordinate information, and $Conv(\cdot)$ refers to the convolution operation with a kernel and stride of 1.

Next, we compute the weights for the boundary transitional regions by calculating the difference between the maximum pooling and average pooling operations (see purple dotted box in Fig. 3). Specifically, we first transform the feature map $F_i^{'}$ into a single-channel activation map F_i^a, which is defined as $F_i^a = SA(F^{N\to1}(F_i^{'}))$. Referring to Eq. (4), the

boundary transitional region $\Psi_\mathbf{i}$ for the feature map F_i^a can be approximated as follows:

$$\Psi_i = \sigma(\sigma(MP(F_i^a)) - \sigma(AP(F_i^a))), \tag{12}$$

where $SA(\cdot)$ denotes the spatial attention operation [Woo *et al.*, 2018], $MP(\cdot)$ denotes the maximum pooling operation, $AP(\cdot)$ denotes the average pooling operation, and σ represents the sigmoid function. Experimental results indicate that a pooling kernel size of 2×2 yields the optimal performance, as detailed in Sec. 4.5.

We then reconstruct the feature map $F_i^{'}$ using the obtained boundary transitional weights Ψ_i. By applying average pooling to filter out noise and using maximum pooling combined with the weights to emphasize important regions, we effectively highlight significant features while minimizing the influence of irrelevant regions. The reconstructed feature map F_i^{re} is defined as follows:

$$F_i^{re} = AP(F_i^{'}) + MP(F_i^{'}) \times \Psi_i. \tag{13}$$

Finally, in order to minimize information loss, we first split the feature map F_i^{re} into two channels (F_i^{re1}, F_i^{re2}) and apply max pooling to preserve critical information in the transitional regions. Concurrently, we use convolution to capture detailed information from other activated regions. The downsampled feature map F_i^{down} is defined as:

$$F_i^{down} = EMA(Conv(F_i^{re1} \times (1 - \Psi_i)) \oplus MP(F_i^{re2} \times \Psi_i)), \tag{14}$$

where $\oplus$ denotes the concatenation operation, and $EMA(\cdot)$ refers to the efficient multi-scale attention mechanism [Ouyang *et al.*, 2023], which helps focus the feature maps on the foreground object, thereby reducing background interference.

The BTRD module enhances the boundary transitional regions containing both strongly and weakly activated features, mitigating the risk of information loss for small objects. This method preserves more unique features of small objects, resulting in multi-scale feature maps with stronger representation capabilities for small objects.

4 Experiments

We have integrated our RAFDE module with the latest YOLO model and conducted experiments on two widely used drone image benchmarks: the **VisDrone dataset** [Du *et al.*, 2019] and the **Drone-vs-Bird dataset** [Coluccia *et al.*, 2021]. The VisDrone dataset consists of 7,019 high-resolution images (2000×1500) containing 10 classes of small, densely packed objects. Of these, 6,471 images are used for training, 548 for validation, and 1,610 for testing. The Drone-vs-Bird dataset includes 1,387 training images and 434 test images, featuring both UAV and environmental data. We evaluated the model's performance using Mean Average Precision (mAP), where mAP_{50} corresponds to an Intersection over Union (IoU) threshold of 0.5, and mAP is calculated as the average across IoU thresholds ranging from 0.5 to 0.95. Due to space limitations, additional experimental results can be found in the supplementary materials.

4.1 Implementation Details

We implemented our RAFDE strategy using PyTorch [Paszke *et al.*, 2019]. All models were trained for 150 epochs, with YOLOv11m serving as the baseline. Our approach employs the same loss function as YOLOv11 [Khanam R, 2024], which includes both object classification loss and bounding box regression loss. For the classification loss, we combine BCELoss [Zheng *et al.*, 2020] and FocalLoss [Li *et al.*, 2020], while for the regression loss, we use CIoULoss [Wang *et al.*, 2023]. The input resolutions were set to 640×640 and 1280×1280 for the VisDrone dataset, and 640×640 for the Drone-vs-Bird dataset. All models were trained using the Adam optimizer with an initial learning rate of 0.01 and a decay rate of 1e-5. Training and testing were conducted on a single RTX A6000 GPU, with batch sizes of 8 and 2 for input resolutions of 640×640 and 1280×1280, respectively.

4.2 Comparison with State-of-the-Art Methods

Table 1 presents a comparison of our RAFDE with state-of-the-art methods on two widely used datasets. On the VisDrone dataset, using an input resolution of 1280×1280, our method was compared with SDPDet [Yin *et al.*, 2024], EFC [Xiao *et al.*, 2024], and YOLOv11 [Khanam R, 2024]. The results demonstrate that our method not only has fewer parameters but also delivers competitive performance, with improvements in mAP (36.80%→38.0%) and mAP_{50} (57.6%→59.6%). Additionally, we evaluated our method on the Drone-vs-Bird dataset using a 640×640 input resolution, where it achieved significant improvements over the baseline, with mAP rising from 52.30% to 54.60% and mAP_{50} increasing from 93.10% to 96.30%. These results demonstrate that our method is highly competitive compared to existing state-of-the-art methods.

4.3 Ablation Studies

We evaluated the effectiveness of each component of our RAFDE method using the VisDrone validation set. YOLOv11m with a 640×640 input resolution was used as the baseline, and the metrics mAP and mAP_{50} were employed for evaluation. As shown in Table 2, the incorporation of the RAFF module, which disentangles co- and uni-activated regions in adjacent low-level and high-level feature maps, effectively mitigates the risk of small object features being overwhelmed during feature fusion. This approach not only reduced the model parameters from 20.06M to 14.40M, but also significantly enhanced its performance, with mAP improving from 26.6% to 28.8%, and mAP_{50} increasing from 43.5% to 47.0%.

By incorporating the BTRD module to enhance the boundary transitional regions, which contain both strongly and weakly activated features, our method effectively reduces the loss of critical features for small objects. This modification led to a further reduction in model parameters from 14.40M to 11.02M, while also achieving a notable performance boost, with mAP improving from 28.8% to 30.0% and mAP_{50} increasing from 47.0% to 48.5%. Compared to the baseline, our RAFDE method significantly improves small object detection performance, with minimal increase in GFLOPs and a substantial reduction in model parameters.

Datasets	Method	Backbone	Size	#P(M)↓	mAP↑	mAP_{50}↑
VisDrone	CZ FCOS Det [Meethal *et al.*, 2023]	ResNet18	1380 × 800	-	33.91	56.20
	GFL V1+CEASC [Du *et al.*, 2023]	ResNet18	1380 × 800	-	28.70	50.70
	FCOS+FGE+SAW [Huang *et al.*, 2024a]	ResNet50	1380 × 800	-	-	51.50
	GFL+EFC [Xiao *et al.*, 2024]	ResNet18	1380 × 800	39.38	30.10	52.10
	YOLOC [Liu *et al.*, 2024a]	ResNet101	1024 × 600	-	29.70	52.40
	SDPDet [Yin *et al.*, 2024]	ResNeXt101	1380 × 800	-	34.20	57.80
	STF-YOLO [Hui *et al.*, 2024]	CSPDarknet53	1280 × 1280	46.74	36.73	-
	YOLOv10 [Wang *et al.*, 2024]	CSPDarknet53	1280 × 1280	16.40	35.60	56.10
	YOLOv11† [Khanam R, 2024]	CSPDarknet53	1280 × 1280	20.06	36.80	57.60
	Brstd [Huang *et al.*, 2024b]	CSPDarknet53	640 × 640	-	27.30	45.9
	SDP [Ma *et al.*, 2023]	CSPDarknet53	1333 × 800	96.70	30.20	52.50
	DINO-DETR [Zhang *et al.*, 2022]	ResNet50	1333 × 800	-	35.80	58.30
	DNTR [Liu *et al.*, 2024b]	ResNet50	1333 × 800	-	33.10	53.80
	RAFDE(ours)	CSPDarknet53	640 × 640	**11.02**	**30.00**	**48.50**
	RAFDE(ours)	CSPDarknet53	1280 × 1280	**11.02**	**38.00**	**59.60**
Drone-vs-Bird	YOLOv5 [Nguyen *et al.*, 2023]	CSPDarknet53	640 × 640	90.96	-	74.60
	DETR+MNMS [Kassab *et al.*, 2024]	ResNet50	640 × 640	-	41.90	82.20
	YOLOv7 [Wang *et al.*, 2023]	CSPDarknet53	640 × 640	37.20	49.20	93.00
	YOLOv10 [Wang *et al.*, 2024]	CSPDarknet53	640 × 640	16.40	50.30	91.70
	YOLOv11† [Khanam R, 2024]	CSPDarknet53	640 × 640	20.06	52.30	93.10
	RAFDE(ours)	CSPDarknet53	640 × 640	**11.02**	**54.60**	**96.30**

Table 1: Performance comparison with state-of-the-art methods on VisDrone and Drone-vs-Bird. The symbol '†' indicates the baseline of RAFDE and the '-' stands for the result that is not reported.

Baseline	RAFF	BTRD	#P(M)	GFLOPs	mAP	mAP_{50}
✓			20.06	68.2	26.6	43.5
✓	✓		14.40	76.4	28.8	47.0
✓	✓	✓	**11.02**	72.0	**30.0**	**48.5**

Table 2: Ablation study of each component on the VisDrone validation set. **RAFF** stands for Regional-Adaptive Feature Fusion module. **BTRD** stands for Boundary Transitional Region-enhanced Downsampling module.

Pooling Kernel Size	**2 × 2**	**3 × 3**	**5 × 5**
mAP	**30.0**	29.1	28.0
mAP_{50}	**48.5**	47.5	45.8

Table 3: Comparing the impact of different pooling kernels on model performance.

4.4 Visualization

Figure 4 demonstrates the effectiveness of our method compared to the baseline on the VisDrone test dataset. Specifically, our approach improves the detection of distant small objects, addressing the issue of missed detections. In the first and second rows, YOLOv11 struggles with detecting distant small objects, missing several distant "people" (see first column, fourth row) and misclassifying "traffic cones" as "people" (see second column, second row). In contrast, our method detects more small objects (see third row, fourth and fifth columns) and reduces false detections (see second column, second and fourth rows). These results highlight the ability of our method to preserve the fine details of small objects, enabling the model to learn from a broader range of small object samples during training, which significantly enhances detection performance.

4.5 Pooling Kernel Size Selection in BTRD

In the BTRD module, a fixed region size is required for scanning in order to extract the maximum and average values within the region. These values are then used to identify the boundary transitional regions containing both strongly and weakly activated features, by calculating their differences. To determine the optimal region size, we tested scanning regions of 2×2, 3×3, and 5×5. The results, shown in Table 3, indicate that larger regions tend to introduce more noise, thereby reducing performance. Therefore, we selected the 2×2 region size for optimal performance.

4.6 Multi-level Semantics Selection in RAFF

When replacing the PAFPN [Liu *et al.*, 2018b] with our RAFF module in YOLOv11, we investigated the influence of low-level detail information and high-level semantic information on the detection of both small and regular-sized objects. As shown in Fig. 1(b), the high-level feature map, used for detecting regular-sized objects, effectively activates with detailed information from adjacent levels. However, as seen in Table 4, the low-level feature map, used for detecting small objects, experiences a significant performance drop when deprived of high-level semantic information. Fig. 5 further demonstrates that the absence of high-level semantic information in the low-level feature map causes small objects to blend into the background, resulting in a considerable loss of detected small objects (highlighted in red boxes in Fig. 5). To address this issue, we directly concatenate high-level and low-level feature maps to retain critical semantic information (as shown in the red box in Fig. 3), leading to a significant improvement in the model's ability to detect small objects.

Figure 4: Comparison of the performance of baseline and our RAFDE on the VisDrone test dataset.

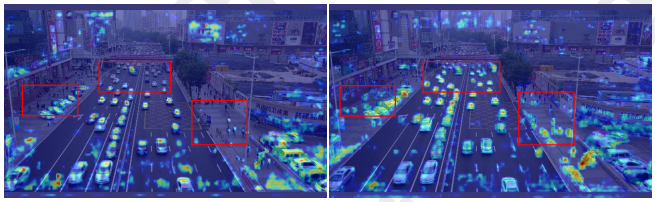

Figure 5: Influence of low-level feature map with or without high-level semantic information on the activation of small objects.

Ours RAFDE	AP	AR	mAP	mAP_{50}
without high-level semantics	52.2	42.4	26.9	43.9
with high-level semantics	**56.7**	**45.4**	**28.8**	**47.0**

Table 4: Effect of low-level feature map with or without high-level semantic information on model performance.

4.7 Effectiveness of RAFDE on Other Backbones

Our plug-and-play method can be easily integrated into any model featuring an FPN structure. To showcase the versatility of RAFDE, we incorporated it into the popular MMDetection framework [Chen *et al.*, 2019] and conducted tests on the VisDrone dataset. Specifically, we selected two two-stage models: Faster R-CNN [Ren *et al.*, 2016] and Cascade R-CNN [Cai and Vasconcelos, 2018]. All experiments were performed on a single RTX A6000 GPU, using a batch size of 4 and an input resolution of 1333 $\times$ 800. The models were trained for 12 epochs with configurations consistent with the baseline. Performance was evaluated using COCO metrics, including AP_s and AP_m, which represent the average precision for small and medium objects, respectively, at IoU thresholds ranging from 0.5 to 0.95. Additionally, AR_s was used to measure the average recall for small and medium objects under the same IoU settings, while AP and AR represent the overall average precision and recall across the full IoU range. Additionally, AP_{50} denotes the average precision at an IoU threshold of 0.5. As shown in the table below, our method **consistently improves model performance across different backbones**, particularly for small objects (e.g., Faster R-CNN: 13.3% $\rightarrow$ 16.5%).

Methods	AP_s	AP_m	AP_{50}	AP	AR_s	AR
Faster-RCNN	13.3	31.7	43.2	26.5	23.4	35.9
+RAFDE(Ours)	**16.5**	**33.3**	**45.3**	**27.7**	**23.5**	**36.5**
Cascade-RCNN	17.1	33.0	44.3	27.9	25.1	37.1
+RAFDE(Ours)	**18.9**	**35.8**	**45.6**	**28.9**	**26.6**	**38.9**

Table 5: Performance of our RAFDE on other backbone networks.

5 Conclusion

In this study, we present a lightweight approach to enhance UAV object detection by investigating the impact of activation strengths of objects at different scales in adjacent levels of feature maps on downsampling and fusion. First, we employed the BTRD module to strengthen the boundary transitional regions, mitigating the loss of crucial boundary features for small objects during downsampling. Second, we introduced the RAFF module, which replaces the conventional FPN structure. This module performs region-adaptive fusion of the co- and uni-activated regions, which are disentangled from adjacent levels, into the current level. This approach reduces feature loss and mitigates the risk of small objects being overwhelmed. Extensive experiments confirm that our RAFDE strategy reduces model parameters while remaining competitive with other state-of-the-art methods.

Acknowledgements

This work is supported in part by the Major Basic Research Project of Shandong Provincial Natural Science Foundation (ZR2024ZD03), National Natural Science Foundation of China (62176141), Taishan Scholar Project of Shandong Province (tsqn202103088), Major science and technology innovation project of Shandong Province (2021CXGC11204), Natural Science Foundation of Shandong Province (ZR202103010201), Shandong Excellent Young Scientists Fund (ZR2024YQ006), Shandong Province Higher Education Institutions Youth Entrepreneurship and Technology Support Program (2023KJ027).

References

[Cai and Vasconcelos, 2018] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[Chen *et al.*, 2019] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[Chen *et al.*, 2023] Tianxiang Chen, Qi Chu, Bin Liu, and Nenghai Yu. Fluid dynamics-inspired network for infrared small target detection. In *IJCAI*, pages 590–598, 2023.

[Coluccia *et al.*, 2021] Angelo Coluccia, Alessio Fascista, Arne Schumann, Lars Sommer, Anastasios Dimou, Dimitrios Zarpalas, Fatih Cagatay Akyon, Ogulcan Eryuksel, Kamil Anil Ozfuttu, Sinan Onur Altinuc, et al. Drone-vs-bird detection challenge at ieee avss2021. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8. IEEE, 2021.

[Dai *et al.*, 2024] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. Freqformer: Frequency-aware transformer for lightweight image super-resolution. *International Joint Conference on Artificial Intelligence*, 2024.

[Du *et al.*, 2019] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[Du *et al.*, 2023] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images supplementary material. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13435–13444, 2023.

[He *et al.*, 2023] Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Wangmeng Xiang, Binghui Chen, Bin Luo, Yifeng Geng, and Xuansong Xie. Damo-streamnet: Optimizing streaming perception in autonomous driving. *International Joint Conference on Artificial Intelligence*, pages 810–818, 2023.

[He *et al.*, 2024] Ang He, Xiaobo Li, Ximei Wu, Chengyue Su, Jing Chen, Sheng Xu, and Xiaobin Guo. Alss-yolo: An adaptive lightweight channel split and shuffling network for tir wildlife detection in uav imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[Hou *et al.*, 2020] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4003–4012, 2020.

[Huang *et al.*, 2024a] Shuqin Huang, Shasha Ren, Wei Wu, and Qiong Liu. Discriminative features enhancement for low-altitude uav object detection. *Pattern Recognition*, 147:110041, 2024.

[Huang *et al.*, 2024b] Sihan Huang, Chuan Lin, Xintong Jiang, and Zhenshen Qu. Brstd: Bio-inspired remote sensing tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Hui *et al.*, 2024] Yanming Hui, Jue Wang, and Bo Li. Stf-yolo: A small target detection algorithm for uav remote sensing images based on improved swintransformer and class weighted classification decoupling head. *Measurement*, 224:113936, 2024.

[Kassab *et al.*, 2024] Mohamad Kassab, Raed Abu Zitar, Frederic Barbaresco, and Amal El Fallah Seghrouchni. Drone detection with improved precision in traditional machine learning and less complexity in single shot detectors. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[Khanam R, 2024] Hussain M Khanam R. Yolov11: An overview of the key architectural enhancements. *arxiv preprint arxiv:2410.17725*, 2024.

[Lei *et al.*, 2023] Tao Lei, Rui Sun, Xuan Wang, Yingbo Wang, Xi He, and Asoke Nandi. Cit-net: Convolutional neural networks hand in hand with vision transformers for medical image segmentation. *International Joint Conference on Artificial Intelligence*, pages 1017–1025, 2023.

[Li *et al.*, 2020] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[Li *et al.*, 2024] Hanqian Li, Ruinan Zhang, Ye Pan, Junchi Ren, and Fei Shen. Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network. *arXiv preprint arXiv:2404.01614*, 2024.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Liu *et al.*, 2018a] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018.

[Liu *et al.*, 2018b] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[Liu *et al.*, 2024a] Chenguang Liu, Guangshuai Gao, Ziyue Huang, Zhenghui Hu, Qingjie Liu, and Yunhong Wang. Yolc: You only look clusters for tiny object detection in aerial images. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Liu *et al.*, 2024b] Hou-I Liu, Yu-Wen Tseng, Kai-Cheng Chang, Pin-Jyun Wang, Hong-Han Shuai, and Wen-Huang Cheng. A denoising fpn with transformer r-cnn for tiny object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Ma *et al.*, 2023] You Ma, Lin Chai, and Lizuo Jin. Scale decoupled pyramid for object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[Meethal *et al.*, 2023] Akhil Meethal, Eric Granger, and Marco Pedersoli. Cascaded zoom-in detector for high resolution aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2046–2055, 2023.

[Moser *et al.*, 2024] Brian B Moser, Arundhati S Shanbhag, Federico Raue, Stanislav Frolov, Sebastian Palacio, and Andreas Dengel. Diffusion models, image super-resolution, and everything: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Nguyen *et al.*, 2023] Duy-Linh Nguyen, Xuan-Thuy Vo, Adri Priadana, and Kang-Hyun Jo. Car detector based on yolov5 for parking management. In *Conference on Information Technology and its Applications*, pages 102–113. Springer, 2023.

[Ouyang *et al.*, 2023] Daliang Ouyang, Su He, Guozhong Zhang, Mingzhu Luo, Huaiyong Guo, Jian Zhan, and Zhijie Huang. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Sun *et al.*, 2022] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. *International Joint Conference on Artificial Intelligence*, pages 1335–1341, 2022.

[Wang *et al.*, 2020] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[Wang *et al.*, 2023] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark 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*, pages 7464–7475, 2023.

[Wang *et al.*, 2024] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Xiao *et al.*, 2024] Yao Xiao, Tingfa Xu, Xin Yu, Yuqiang Fang, and Jianan Li. A lightweight fusion strategy with enhanced inter-layer feature correlation for small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Yang *et al.*, 2024] Guoyu Yang, Jie Lei, Hao Tian, Zunlei Feng, and Ronghua Liang. Asymptotic feature pyramid network for labeling pixels and regions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Yin *et al.*, 2024] Nengzhong Yin, Chengxu Liu, Ruhao Tian, and Xueming Qian. Sdpdet: Learning scale-separated dynamic proposals for end-to-end drone-view detection. *IEEE Transactions on Multimedia*, 2024.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[Zheng *et al.*, 2020] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.