Perception and Regression Network for Small Object Detection in UAV Aerial Photography

1st Xiaofeng Wang
School of Cyber Security and Computer
Hebei University
Baoding, China
20247019051@stumail.hbu.edu.cn

2nd Wenzhu Yang*
School of Cyber Security and Computer
Hebei University
Machine Version Engineering Research Center
Hebei University
Baoding, China
wenzhuyang@hbu.edu.cn

***Abstract*—Unmanned Aerial Vehicles (UAVs) object detection has widespread applications in the real world. However, due to the small size, dense distribution and blurred features of UAV images, traditional object detection algorithms encounter significant challenges. To address these issues, this paper proposes a Perception and Regression Network (PRNet) built upon YOLOv11, which incorporates multi-scale feature fusion, spatial attention mechanisms, and regression optimization. First, we design the C3k2-MSF module, which enhances the model's ability to perceive objects at multiple scales. Second, we design the BG module that integrates the Bidirectional Feature Pyramid Network (BiFPN) module and the Global-to-Local Spatial Aggregation (GLSA) module, aiming to improve small object detection accuracy. Finally, we propose Coord-Dyhead, an advanced detection head that integrates our newly designed Coord-SE module into the Dynamic Head (Dyhead) architecture. Experimental results indicate that our PRNet achieves significant improvements of 10.8% and 7.4% in mAP over the baseline model on the VisDrone2019 and AI-TOD datasets respectively, while maintaining comparable model parameters.**

***Index Terms*—Unmanned aerial vehicles, Contextual information learning, Multi-scale feature fusion, Channel attention mechanism**

I. Introduction

Object detection in Unmanned Aerial Vehicle (UAV) imagery has emerged as a critical technology for identifying and localizing targets using on-board sensors. Due to their exceptional maneuverability, high real-time responsiveness, and low operational costs, UAVs are extensively deployed as a standard tool for image collection. Therefore, drone aerial images are widely applied in intelligent transportation systems, precision agriculture, and disaster relief operations. While UAV imagery offers the advantage of wide-area coverage, it presents unique challenges compared to ground-based photography. Specifically, it has relatively low resolution quality, small detection object sizes (less than 32×32 pixels [1]), dense distribution, and instance overlap issues. Furthermore, UAV imagery is often affected by environmental factors, including atmospheric turbulence and illumination variations, which adversely impact object detection performance. These constraints underscore the urgent need for specialized refinements in existing detection methodologies.

Deep learning-based object detectors generally fall into two categories: two-stage and single-stage. Two-stage object detectors, such as Faster RCNN [2], follow an implementation process that first extracts candidate regions and then classifies and identifies these regions. However, two-stage object detectors are computationally intensive and exhibit slower detection speeds. Conversely, single-stage detectors, represented by SSD [3] and the YOLO series [4] - [11], perform direct regression and classification, offering a superior trade-off between speed and accuracy. Notably, YOLOv11 [11] achieves impressive performance through architectural refinements and GPU-specific optimizations, maintaining high processing speeds with comparable parameter efficiency. Nevertheless, standard YOLO architectures exhibit limitations when confronted with the drastic scale disparities inherent in drone aerial views. This limitation largely stems from their reliance on generic feature fusion mechanisms. While effective for general detection tasks, standard multi-scale fusion strategies often fail to preserve the fine-grained features essential for small objects in UAV imagery, or conversely, they introduce excessive computational overhead incompatible with edge deployment.

To tackle these challenges, we design a Perception and Regression Network (PRNet), which builds upon YOLOv11m. In this model, we designed the C3k2-MSF module in the backbone network to strengthen the model's contextual perception capacity. In the neck network, we introduced and integrated the Bidirectional Feature Pyramid Network (BiFPN) [12] and Global-to-Local Spatial Aggregation (GLSA) [13] modules to obtain the BG module, which boosts the discriminative power of features pertaining to small targets. Finally, we design Coord-DyHead by integrating a Coordinate-aware SE module into the Dynamic Head (Dyhead) [14], enabling the network to simultaneously suppress redundant features and maintain high sensitivity to target positions.

This study's key contributions are systematically outlined below:

1. We propose the Multi Scale Fusion (MSF) module and incorporated it into the C3k2 structure (C3k2-MSF), augment-

* Corresponding author.

ing the model's capability to sense multi-scale object, with particularly outstanding performance in small target detection.

2. We design the BG module by integrating BiFPN into the neck network to facilitate feature fusion, and integrated it with GLSA module to increase global and local spatial attention, thus elevating the model's sensitivity to small-scale object.

3. We propose the Coord-DyHead, which integrates a coordinate-aware attention mechanism into the DyHead. By replacing standard global pooling with direction-aware feature aggregation, this module effectively filters background noise while preserving critical spatial information for small targets, enhancing the model's localization precision.

Due to ongoing intellectual property protection, the source code is available from the corresponding author upon reasonable request.

II. Related Work

A. Contextual modeling for small object detection

Contextual information, which encodes the semantic and spatial relationships between an object and its surrounding environment, plays a pivotal role in detecting small targets with limited visual cues. Traditional CNN-based approaches have extensively explored mechanisms to enhance context perception. As representative works, DSSD [15] introduces deconvolution layers and skip connections to infuse deep semantic context into shallow feature maps. To mitigate feature erosion during downsampling, DenseNet [16] adopts a dense connectivity pattern, ensuring maximum information flow between layers. Similarly, PyramidBox [17] utilizes a context-sensitive prediction module to learn semantic feature distributions from anchor priors. Furthermore, FA-SSD [18] explicitly integrate features from the target's periphery, thereby enriching the feature representation for small-scale crowd detection.

Despite these advancements, inherent limitations remain in pure CNN architectures. Convolutional operations are essentially local, processing information within a fixed receptive field. Consequently, they struggle to model long-range dependencies and effectively capture the global relationship between distant pixels, which is indispensable for distinguishing small UAV targets from complex backgrounds where the target's intrinsic features are scarce. While recent approaches, exemplified by TPH-YOLOv5 [6], have attempted to incorporate attention modules to boost performance, few have effectively integrated the Transformer encoder into modern CSP-based architectures, such as the C3k2 design in YOLO11, to balance efficiency and global context modeling in UAV tasks. To address this, we propose the C3k2-MSF module, which leverages the self-attention mechanism to capture global contextual information, thereby significantly enhancing the model's perception capabilities for tiny targets.

B. Multi-scale feature fusion

Integrating features across varying resolutions is fundamental for generating scale-invariant representations, particularly for detecting small objects where visual cues are scarce.

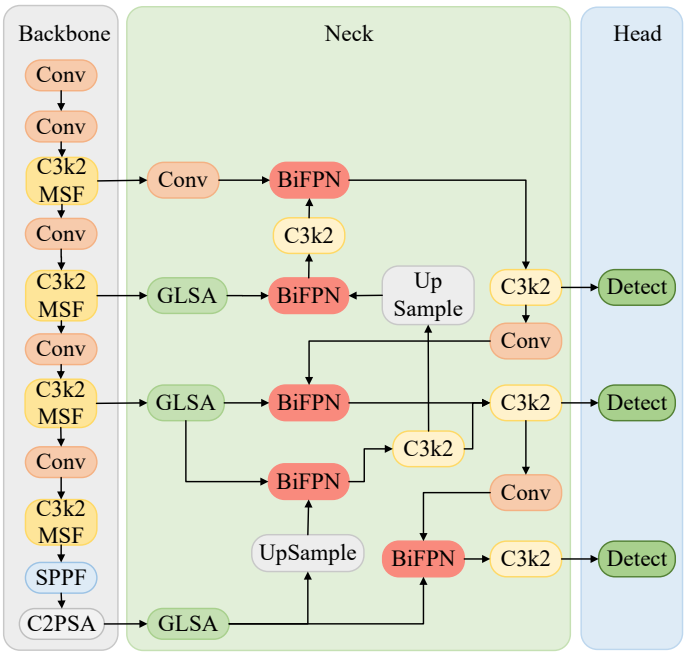


Fig. 1. The network structure of the PRNet, mainly consists of the C3k2-MSF module, BiFPN module, GLSA module and Coord-Dyhead. The SPPF module, C2PSA module are taken from YOLOv11.

While early single-scale architectures suffered from inevitable information loss, modern approaches leverage multi-scale hierarchies to retain both fine-grained spatial details and coarse semantic information. However, simply aggregating multi-scale features often introduces computational redundancy and semantic misalignment due to the discrepancy between layers.

To mitigate these issues, significant efforts have been devoted to optimizing information flow. PANet [19] pioneered a bottom-up path augmentation to shorten the information path between lower layers and topmost features. Building on this, Aug-FPN [20] addresses the semantic gap before fusion through residual feature augmentation and soft region proposal selection. Similarly, PRB-FPN [21] further enhance cross-scale detection by dynamically adjusting receptive fields and employing hierarchical designs to better perceive targets of varying sizes.

Despite these advancements, standard fusion mechanisms often treat features from different scales equally, ignoring their unequal contribution to the detection task. To address this, we adopt BiFPN, which introduces learnable weights to dynamically prioritize more informative features during fusion. Furthermore, to compensate for the potential loss of spatial details during pyramid pooling, we integrate GLSA into the fusion architecture. By synergizing BiFPN's weighted fusion with GLSA's global-local spatial attention, our approach effectively highlights small target details while suppressing background noise.

C. Object detection head design

The fundamental pursuit of object detection is to simultaneously determine "what" an object is (classification) and

”where” it is located (localization). Consequently, the design of the detection head implies a need to address three critical challenges: scale awareness for varying object sizes, spatial awareness for geometric deformations, and task awareness to align the distinct objectives of classification and regression. While the Dynamic Head framework unifies these dimensions, its reliance on global pooling operations risks discarding high-frequency spatial cues, which are vital for localizing small targets.

To bridge this gap, we introduce Coord-DyHead, an enhanced architecture that integrates Coordinate-aware Attention into the dynamic fusion process. Unlike standard Squeeze-and-Excitation (SE) [22] modules that indiscriminately compress spatial dimensions into a single scalar, Coord-DyHead decomposes attention into two parallel 1D feature encoding processes. By aggregating features along the horizontal and vertical directions separately, our module explicitly encodes precise positional information into the channel recalibration process. This results in a dual-adaptive mechanism that not only suppresses background noise but also preserves the spatial coordinates of targets, significantly boosting the model's discriminative power in complex scenarios.

III. Methods

Adhering to the classic YOLO architecture, PRNet is structured into three integrated components, namely the Backbone, Neck, and Head, as illustrated in Fig. 1. The backbone utilizes an optimized CSPDarkNet [8] equipped with our novel C3k2-MSF modules to enhance multi-scale feature extraction capabilities beyond standard computational limits. For the neck, we reformulate the feature fusion mechanism by replacing the conventional FPN-PANet with a BG structure, which incorporates BiFPN and GLSA modules to robustly propagate semantic and spatial information. At the detection end, we propose Coord-DyHead, a coordinate-aware dynamic head that unifies scale, spatial, and task representations. This specific design suppresses redundant features while sharpening localization accuracy, making PRNet particularly effective for identifying small targets in high-resolution UAV imagery.

A. C3k2-MSF

To surmount the limitations of fixed receptive fields in the standard C3k2 module of YOLOv11, which often struggles to balance global semantic context with local spatial details, we introduce the C3k2-MSF module. As a structural evolution, C3k2-MSF replaces the conventional bottleneck stack with a series of C3k-MSF units, as illustrated in Fig. 2. Inspired by the meta-architecture of Transformer encoders [23], the C3k-MSF unit is engineered to capture long-range dependencies while maintaining the parallel computing efficiency of CNNs. By embedding our novel Multi-Scale Fusion (MSF) mechanism within this efficient backbone, the module explicitly enhances the network's capacity to perceive targets across varying scales.

As illustrated in Fig. 3, the core innovation lies within the MSF module, which decouples contextual abstraction from detail extraction through a multi-branch topology. To address the challenge of scale variation, the input features are processed via heterogeneous kernel branches. Specifically, large-kernel convolutions are employed to expand the effective receptive field, capturing broad contextual information essential for understanding the scene. Complementing this, a dedicated Detail Extractor branch is designed to prevent the erosion of fine-grained features, which is particularly critical for small object detection. This branch utilizes an inverted bottleneck structure composed of a 1×1 projection, a 3×3 depth-wise convolution, and a 1×1 expansion to efficiently encode high-frequency spatial cues and boundary information with minimal computational cost. To synthesize these diverse representations, the MSF module moves beyond simple element-wise summation by integrating a Channel Attention mechanism. This component serves as an adaptive gatekeeper, dynamically recalibrating the channel weights to emphasize discriminative features while suppressing background noise.

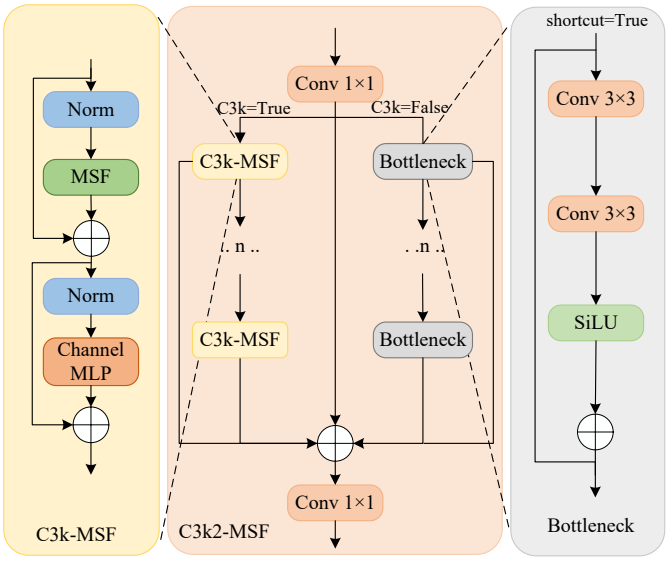


Fig. 2. Structure of the C3k2-MSF modules. The C3k2-MSF module embeds the MSF module into the C3k2 framework to strengthen the model's capacity for contextual information perception and feature extraction.

Unlike the original C3k2 module, which treats all spatial locations and channels equally, our design ensures that the network focuses on the most informative scales for the target at hand. The fused features are finally projected through a linear layer, resulting in a representation that is both semantically rich and spatially precise. This holistic design allows the C3k2-MSF to significantly outperform the baseline in localization accuracy without sacrificing inference speed.

B. BG module

The standard YOLOv11 neck employs a PANet structure to facilitate multi-scale feature interaction, as illustrated in Fig. 4. While effective for general object detection, it relies on simple concatenation and element-wise addition, which treats all features equally. This approach often fails to prioritize the faint, fine-grained details critical for detecting small objects in drone

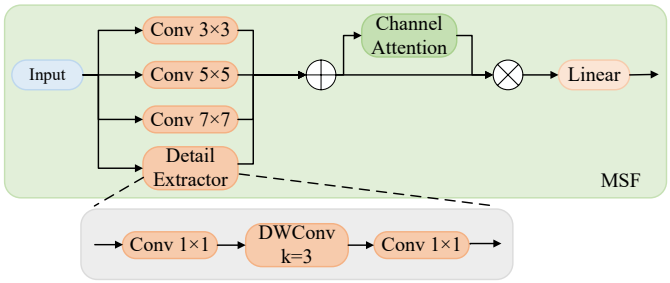


Fig. 3. Structures of MSF module and Detail Extractor. The MSF module contain three convolution kernels of different sizes and a Detail Extractor. The MSF module integrates Detail Extractor to reinforce the network's capability to capture small target features.

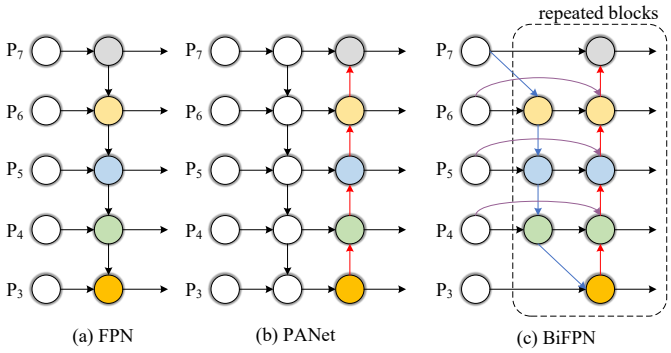


Fig. 4. (a) FPN introduces a top-down pathway to fuse multi-scale features; (b) PANet adds an additional bottom-up pathway on top of FPN; (c) BiFPN removes nodes with only one incoming edge and introduces multiple skip connections.

imagery. Although BiFPN improves scale-invariance through learnable weights and cross-scale connections, it primarily optimizes channel-wise fusion and lacks explicit mechanisms to model spatial relationships. Consequently, minute targets are prone to being submerged by background noise during deep feature transmission.

To bridge this gap, we introduce the BG module, a hybrid feature fusion network that synergistically integrates the structural efficiency of BiFPN with the spatial sensitivity of the Global-Local Spatial Attention (GLSA) mechanism. Instead of a naive replacement, we strategically embed the GLSA module into the critical feature fusion nodes (specifically layers 11-13) of the BiFPN pathway. This topological integration allows the network to leverage BiFPN for efficient cross-scale information flow while utilizing GLSA to explicitly refine spatial features at each fusion stage, ensuring that small target details are preserved and enhanced against complex backgrounds.

The core of our BG design is the GLSA module, which we adopt to decouple spatial modeling into two complementary branches: Global Spatial Attention (GSA) and Local Spatial Attention (LSA). To balance performance with computational overhead, we employ a channel-split strategy. As shown in Fig. 5, the input feature map $\mathcal{F}_{in}$ is split along the channel dimension into a global branch $\mathcal{F}_g$ and a local branch $\mathcal{F}_l$. These are processed in parallel and then fused:

$$\mathcal{F}_{\text{out}} = \text{Conv}_{1\times1}\left(\text{Concat}\left(\text{GSA}(\mathcal{F}_g), \text{LSA}(\mathcal{F}_l)\right)\right) \quad (1)$$

where GSA and LSA denote the transformations of the two branches, respectively.

1) Global context modeling (GSA): In UAV scenarios, distinguishing targets from similar environmental textures requires global context. The GSA branch captures long-range dependencies. It generates a global attention map via a Softmax operation to re-weight the features:

$$GSA(\mathcal{F}_g) = \text{MLP}\left(\text{Softmax}\left(\mathcal{F}_g^T \mathbf{W}_q\right) \cdot \mathcal{F}_g\right) + \mathcal{F}_g \quad (2)$$

This mechanism ensures that the representation of any specific pixel is informed by the entire scene, effectively suppressing false positives.

2) Local detail enhancement (LSA): Complementing the global view, the LSA branch is specifically tailored to prevent the erosion of small object features. It employs a bottleneck structure with cascaded depth-wise convolutions to extract high-frequency local cues. The local attention is applied via a gating mechanism:

$$LSA(\mathcal{F}_l) = \sigma\left(\mathcal{B}_{local}(\mathcal{F}_l)\right) \odot \mathcal{F}_l + \mathcal{F}_l \quad (3)$$

where $\mathcal{B}_{local}$ represents the local convolution block and σ is the sigmoid function. By focusing on local neighborhoods, LSA significantly improves boundary localization for small targets.

By unifying these components, the BG module transforms the feature fusion process from a passive combination into an active, context-aware recalibration, significantly boosting detection accuracy for small targets with minimal latency overhead.

C. Coord-Dyhead

Standard detection heads often struggle to balance the conflicting demands of classification and localization, specifically in complex drone imagery. While the Dynamic Head framework unifies scale, spatial, and task awareness into a coherent representation, its efficacy is compromised by the use of Global Average Pooling (GAP). GAP aggressively compresses spatial signals into scalars, inevitably discarding positional cues essential for localizing small targets. To rectify this loss of spatial fidelity, we propose Coord-DyHead, which strategically integrates our designed Coord-SE module into the unified attention gates. Unlike reductive GAP operations, the Coord-SE module decomposes channel attention into two parallel 1D feature encoding processes, allowing the network to capture long-range dependencies along one direction while preserving precise positional information along the other. This facilitates more accurate object localization even during dynamic scale and task reconfiguration.

As illustrated in Fig. 6, the Coord-SE module preserves spatial fidelity by decomposing channel attention into two parallel 1D coordinate encoding processes. Instead of using global pooling, it aggregates features along the horizontal and vertical directions to generate direction-aware attention maps,

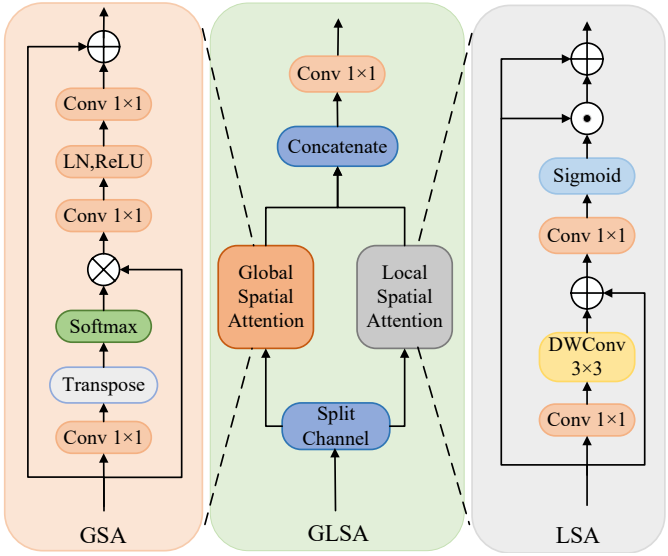


Fig. 5. The structure of Global-to-Local Spatial Aggregation (GLSA) module, it is composed of Global Spatial Attention (GSA) and Local Spatial Attention (LSA).

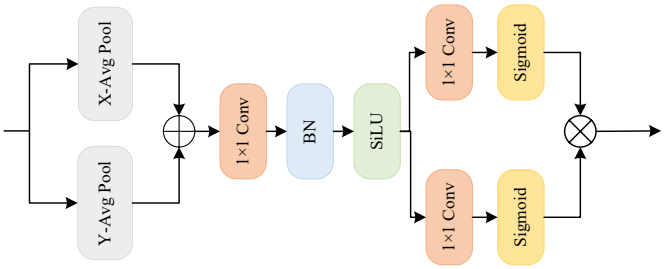


Fig. 6. Structure of Coord-SE Block

which are then applied to the input feature map via element-wise multiplication. This mechanism allows the network to capture long-range dependencies while retaining precise positional cues essential for accurate localization.

Coord-DyHead treats the detection head input as a 3D tensor $\mathcal{F} \in \mathbb{R}^{L\times S\times C}$ and applies a sequential attention mechanism:

$$W(\mathcal{F}) = \pi_C\left(\pi_S\left(\pi_L(\mathcal{F})\cdot\mathcal{F}\right)\cdot\mathcal{F}\right)\cdot\mathcal{F} \tag{4}$$

where $\pi_L(\cdot)$, $\pi_S(\cdot)$, and $\pi_C(\cdot)$ denote scale-aware, spatial-aware, and task-aware attention functions, respectively.

1) Coordinate-Guided Scale Awareness (π_L): Instead of relying on global statistics to weight different feature levels, we introduce coordinate-aware pooling to preserve spatial sensitivity during scale fusion. Specifically, standard global average pooling is decomposed into two one-dimensional pooling operations along the horizontal and vertical directions, enabling the aggregation of direction-aware contextual information. Based on these spatially preserved descriptors, the scale fusion weights are generated as

$$\pi_L(\mathcal{F})\cdot\mathcal{F} = \sigma\left(\Phi_{Coord}\left(f\left(\frac{1}{SC}\sum_{S,C}\mathcal{F}\right)\right)\right)\cdot\mathcal{F} \tag{5}$$

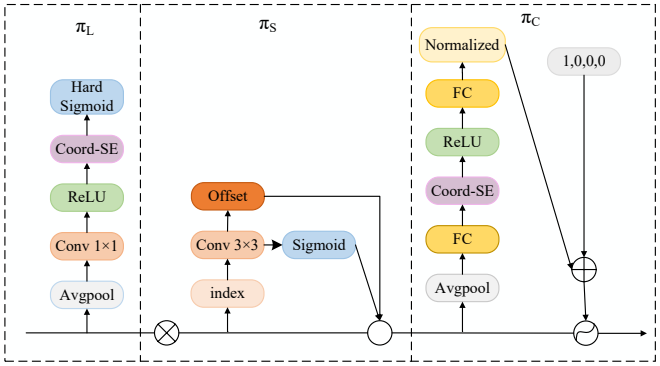


Fig. 7. Structure of Coord-Dyhead

where Φ_{Coord} represents the coordinate attention operation that aggregates vertical and horizontal contexts. This ensures that the dynamic fusion of multi-scale features ($\mathcal{F}$) is guided by spatially preserved semantic descriptors rather than mere global statistics.

2) Spatial-Aware Attention (π_S): Following the sparse sampling paradigm, π_S focuses on reinforcing discriminative spatial regions. It utilizes deformable convolutions to aggregate features across levels at aligned spatial locations:

$$\pi_S(\mathcal{F})\cdot\mathcal{F} = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k}\cdot\mathcal{F}(l;p_k+\Delta p_k;c)\cdot\Delta m_k \tag{6}$$

where L denotes the number of feature levels, K is the number of sampled locations per level, and $w_{l,k}$ represents the learnable sampling weight at the k-th location of the l-th level. Δp_k is the learned offset that enables adaptive spatial sampling, ensuring the module focuses on the most relevant object parts, while Δm_k denotes the modulation scalar that controls the contribution of each sampled feature.

3) Coordinate-Guided Task Awareness (π_C): The final stage dynamically tunes feature channels for classification or regression. We replace the standard dynamic activation with a coordinate-sensitive gating mechanism. This allows the network to activate channels based on both target category and spatial position:

$$\pi_C(\mathcal{F})\cdot\mathcal{F} = \max\left(\Psi_{Coord}(\alpha^1\mathcal{F}+\beta^1), \Psi_{Coord}(\alpha^2\mathcal{F}+\beta^2)\right) \tag{7}$$

where Ψ_{Coord} applies coordinate attention to the task-specific projections. $\{\alpha^i, \beta^i\}$ denote learnable affine transformation parameters for different tasks (classification and regression), and the channel-wise max operation selects the dominant task-aware response for each feature channel.

By embedding Coordinate Attention into the Dynamic Head, Coord-DyHead addresses the critical loss of spatial information caused by global pooling in small object detection. This integration allows the network to capture long-range dependencies while preserving precise positional cues, enabling robust localization of minute targets against complex drone backgrounds without increasing computational overhead.

IV. Experimental results

A. Datasets

(1) VisDrone2019 [24] is a dataset consisting of 10,209 static images, recorded by a variety of UAV-mounted cameras and encompassing an extensive variety of scenarios with variations in environment, location, object types, and density. Divided into 6,471 for training, 548 for validation, and 3,190 for testing, these images contain over 2.6 million bounding boxes or point annotations that label objects of interest. Including 10 object categories, the dataset features image resolutions ranging from 540 × 960 to 2000 × 1500 pixels.

(2) AI-TOD [25] is a dedicated dataset consisting of 28,036 aerial images for tiny object detection within aerial photographs. Distinguished from current remote sensory object detection datasets, notably smaller than objects in other public datasets, this dataset contains objects with an average size of approximately 12.8 pixels. The dataset containing a total of 700,621 object instances categorized into eight classes. Following official specifications, we use 14,018 images (11,214 training, 2,804 validation) for training, with performance evaluated on the 14,018-image test set.

B. Metrics

Following standard COCO evaluation protocols [26], we employ Average Precision (AP) as the primary metric. Specifically, we evaluate detection performance using mAP (averaged over IoU thresholds from 0.5 to 0.95), $mAP_{0.5}$, and $mAP_{0.75}$. In addition, we report AP for small (AP_s), medium (AP_m), and large (AP_l) objects to provide a comprehensive assessment.

C. Experimental setup

All experiments in this study were conducted on a workstation equipped with an NVIDIA 4090 GPU. The implementation leverages the PyTorch 2.3.0 framework, with Python 3.10 and CUDA 11.3 as the underlying environments. For parameter optimization, the SGD optimizer was adopted, configured with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Training was carried out with a batch size of 16, input images resized to 640 × 640, and a total of 200 epochs.

D. Comparisons with previous methods

To comprehensively evaluate the effectiveness of our proposed PRNet, we benchmarked it against a diverse set of representative object detection methods.

Table I and Table II present the comparative results on the VisDrone2019 test set and AI-TOD validation set, respectively. While two-stage detectors like Cascade R-CNN achieve decent accuracy, they suffer from excessive parameter counts and high latency. Conversely, anchor-free methods and standard YOLO models often struggle with minute targets. This is primarily because their aggressive downsampling strategies and global pooling operations lead to the loss of fine-grained spatial information essential for small object localization. For instance, although YOLOv10 improves speed by removing NMS, its accuracy drops due to the lack of specialized handling for dense, tiny objects. Specifically, our proposed method significantly outperforms baselines, achieving the best results in mAP and $mAP_{0.5}$ while surpassing the latest YOLOv11m by 8.2% and 3.7% in the critical AP_s metric on the VisDrone2019 and AI-TOD datasets, respectively. In addition, PRNet's FPS has increased by 5% compared to the baseline model, indicating an improvement in the model's detection speed.

As shown in Table III, we conducted comparative experiments against other advanced modules on the VisDrone2019 dataset to validate the design of our key components, namely C3k2-MSF, BG, and Coord-DyHead.

We compared our module against lightweight and dynamic alternatives, including C3k2-Faster, C3k2-iRMB, and C3k2-SCConv. While methods like C3k2-iRMB introduce dynamic fusion, they fail to fully capture the multi-scale context required for small targets. Our C3k2-MSF demonstrates superior feature extraction capability by balancing local and global context.

In the Neck, our BG module outperforms the standalone implementations of both BiFPN and GLSA. By synergizing the multi-scale feature fusion of BiFPN with the global-local spatial attention of GLSA, the proposed module achieves a performance level that neither component could reach in isolation. This proves that integrating these two mechanisms is essential for mitigating localization errors in complex backgrounds.

Similarly, for the detection head, while Dynamic Head variants (LQEHead, LADH-Head) offer efficiency, they lack the specific coordinate-sensitivity of our Coord-Dyhead. Consequently, our method achieves the highest accuracy, proving that our tailored improvements are more effective for UAV-based detection than generic SOTA modules.

In summary, our improved PRNet has achieved excellent detection results for small target detection, and the model exhibits strong generalizability and is capable of adapting to diverse datasets.

E. Ablation experimental result

As detailed in Table IV, we conducted an ablation study to quantify the contribution of each component. The results demonstrate that each proposed strategy confers a distinct performance boost to the YOLOv11m baseline. Specifically, incorporating the C3k2-MSF module yields a substantial 3.6% increase in mAP. Most notably, the BG module, which synergizes BiFPN with GLSA, delivers the largest gain of 6.2% mAP, primarily owing to its superior multi-scale feature fusion and enhanced small-target localization capability. Furthermore, the introduction of Coord-Dyhead improves mAP by 1.5%, attributing to its ability to suppress background noise while adaptively highlighting salient features across different scales. Crucially, the progressive integration of these modules leads to cumulative gains, with the full configuration achieving the highest detection accuracy.

TABLE I
PERFORMANCE COMPARISON OF VARIOUS MODELS ON THE VISDRONE2019 DATASET, WITH THE OPTIMAL RESULTS HIGHLIGHTED IN BOLDFACE.

Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$	AP_s	AP_m	AP_l	Params(M)	FLOPs(G)	FPS
Faster-RCNN [2]	19.4	32.9	19	9.5	30.9	42.9	41.1	208	40.2
Cascade-RCNN [27]	19.7	32.6	21	9.9	30.9	40.6	69.3	236	16.5
RetinaNet [28]	16.4	27.6	15.5	5.5	25.1	42.7	36.5	210	28.4
CenterNet [29]	14.8	28.4	14.1	6.3	24.3	32.3	32.7	70.2	-
ATSS [30]	17.9	31	18.5	8.5	27.9	42.3	31	**49.8**	-
YOLOv5m [6]	18.8	32.8	19.1	8.9	29.4	40.5	25	64	103.9
YOLOv8m [8]	19.7	34.1	20	9.4	30.7	44	25.9	78.7	**113.3**
YOLOv9m [9]	19.3	33.6	19.4	8.9	30.3	43.2	20	76.5	52.8
YOLOv10m [10]	18.9	33	19.2	8.9	29.2	40	**16.5**	63.5	97.7
YOLOv11m [11]	19.5	34.2	19.8	9.7	29.8	40.2	20.1	68	91.9
TPH-YOLOv5 [6]	19.8	34.2	20.3	8.7	28.8	42.4	37	53.5	57.6
Ours	**21.6**	**37**	**22.2**	**10.5**	**33.1**	**43.8**	19.8	76.1	96.5

TABLE II
PERFORMANCE COMPARISON OF VARIOUS MODELS ON THE AI-TOD DATASET, WITH THE OPTIMAL RESULTS HIGHLIGHTED IN BOLDFACE.

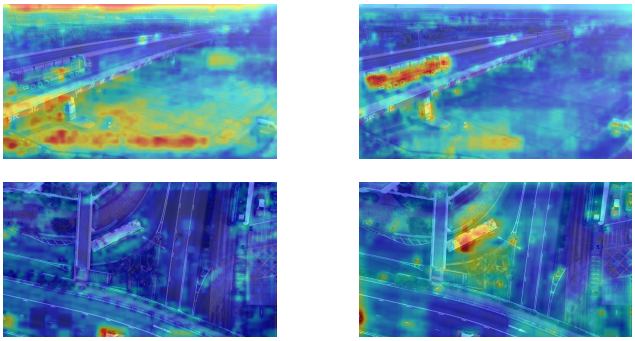
Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$	AP_s	AP_m
Faster-RCNN [2]	9.5	27.1	4	8.8	21.4
Cascade-RCNN [27]	10.9	29.2	5.5	10.5	25.2
RetinaNet [28]	5.2	15.1	2.2	4.8	14.1
CenterNet [29]	5.8	16.2	2.7	5.7	9.2
ATSS [30]	12.8	30.6	8.5	11.6	19.5
YOLOv5m [6]	14.2	32.6	9.8	13.3	33.3
YOLOv8m [8]	14.9	34.3	10.2	13.9	35.8
YOLOv9m [9]	15.8	36.8	11.3	15.1	35.1
YOLOv10m [10]	15.2	33.8	11.6	14.7	31.4
YOLOv11m [11]	16.3	37	11.5	16.1	35.2
TPH-YOLOv5 [6]	16.5	36.6	11.7	15.4	34.8
Ours	**17.5**	**40.8**	**12.5**	**16.7**	**36.2**

TABLE III
COMPARATIVE EXPERIMENTS WITH DIFFERENT C3K2 IMPROVEMENTS, NECK IMPROVEMENTS, AND HEAD IMPROVEMENTS ON THE VISDRONE2019 DATASET, WITH THE OPTIMAL RESULTS HIGHLIGHTED IN BOLDFACE.

Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$	AP_s	AP_m
C3k2-faster [31]	19.2	33.5	19.6	9.3	29.6
C3k2-iRMB [32]	18.7	32.9	18.9	8.9	28.8
C3k2-SCConv [33]	19.3	33.6	19.7	9.2	30
C3k2-MSF	**20.2**	**35.2**	**20.7**	**10.1**	**31.1**
BiFPN [12]	20.1	34.7	20.7	9.7	31.4
GLSA [13]	19.6	34.2	19.9	9.6	30.2
BG	**20.7**	**35.9**	**21.3**	**10.1**	**32.1**
EfficientHead [12]	19.4	33.8	19.8	9.4	29.8
LQEHead [34]	19.5	33.9	20	9.5	30.1
SEAMHead [35]	19.6	34.1	19.9	9.6	30.1
LADH-Head [36]	19.4	33.7	19.7	9.3	29.7
Coord-Dyhead	**19.8**	**34.1**	**20.4**	**9.7**	**30.7**

TABLE IV
RESULTS OF THE ABLATION EXPERIMENT ON THE VISDRONE2019 DATASET, WITH THE OPTIMAL RESULTS HIGHLIGHTED IN BOLDFACE.

Method	mAP	$mAP_{0.5}$	$mAP_{0.75}$	AP_s	AP_m
-	19.5	34.2	19.8	9.7	29.8
+C3k2-MSF	20.2	35.2	20.7	10.1	31.1
+BG	20.7	35.9	21	21.3	10.1
+Coord-Dyhead	19.8	34.1	20.4	9.7	30.7
C3k2-MSF+BG	20.9	36	21.4	10.2	32.5
all	**21.6**	**37**	**22.2**	**10.5**	**33.1**



(a) (b)

Fig. 8. Heatmaps of object detection result. (a) YOLOv11m. (b) PRNet

F. Visual evaluation

For a more intuitive comparison between our algorithm and baseline methods, we visualize the class activation heatmaps generated by Grad-CAM [37] on the VisDrone2019 test set, as shown in Fig. 8. It can be observed from these figures that the attention of the baseline model is severely deviated. For instance, it fails to focus on the person under the overpass, whereas PRNet successfully pays good attention to the small targets in the image. Experimental results demonstrate that PRNet exhibits superior robustness, thereby validating the efficacy of the improvements we propose.

V. CONCLUSION

In this paper, we introduced PRNet, an enhanced object detection framework tailored for the challenges of drone-captured imagery. By integrating three strategic innovations into the YOLOv11m baseline, namely C3k2-MSF for contextual perception, the BG module for robust global–local feature fusion, and Coord-DyHead for scale-aware detection, our method achieves a superior trade-off between accuracy and efficiency. Extensive experiments on the VisDrone2019 and AI-TOD datasets demonstrate that PRNet significantly outperforms state-of-the-art methods, particularly in small object detection, while incurring negligible computational overhead.

While PRNet exhibits promising results, there is potential for further optimization in the backbone architecture to handle extreme environmental complexities. Future research will

focus on exploring lightweight feature fusion mechanisms to further reduce latency for deployment on resource-constrained UAV hardware. We hope PRNet serves as a solid baseline for future advancements in aerial object detection.

ACKNOWLEDGMENT

The author thank the National Science Foundation of Hebei Province (F2024201012) for their financial support and the support of the High-Performance Computing Center of Hebei University.

REFERENCES

[1] T. Zhang, G. Wang, Y. Zhuang, H. Chen, H. Shi and L. Chen, "Feature enhanced CenterNet for object detection in remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2020, pp. 1639-1642.

[2] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137-1149, 2017.

[3] W. Liu, D. Anguelov, D. Erhan, et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21-37.

[4] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779-788.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6517-6525.

[6] X. Zhu, S. Lyu, X. Wang and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on Transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2021, pp. 2778-2788.

[7] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7464-7475.

[8] M. Yaseen, "What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector," *arXiv preprint arXiv:2408.15857*, 2024.

[9] C.-Y. Wang, I.-H. Yeh and H.-Y. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[10] A. Wang, H. Chen, L. Liu, et al., "YOLOv10: real-time end-to-end object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024, pp. 107984-108011.

[11] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[12] M. Tan, R. Pang and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10778-10787.

[13] F. Tang, Z. Xu, Q. Wang, J. Hou, X. Su and J. Liu, "DuAT: Dual-aggregation transformer network for medical image segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, 2023, pp. 343-356.

[14] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu and L. Yuan, "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 7369-7378.

[15] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi and A. C. Berg, "DSSD: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[16] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700-4708.

[17] T. Xu, D. K. Du, Z. He and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 797-813.

[18] Y. Zheng, Y. Wang, F. Li and Y. Zhu, "FA-SSD: A small object detection algorithm based on feature alignment," in *Proc. Int. Conf. Pattern Recognit. Artif. Intell. (PRAI)*, 2022, pp. 375-380.

[19] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8759-8768.

[20] C. Guo, F. Bin, Q. Zhang, S. Xiang and C. Pan, "AugFPN: Improving multi-scale feature learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12595-12604.

[21] P. -Y. Chen, M. -C. Chang, J. -W. Hsieh and Y. -S. Chen, "Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection," in *IEEE Trans. Image Process.*, vol. 30, pp. 9099-9111, 2021.

[22] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132-7141.

[23] A. Vaswani, N. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[24] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin and Q. Hu, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, 2019, pp. 213-226.

[25] J. Wang, W. Yang, H. Guo, R. Zhang and G. -S. Xia, "Tiny object detection in aerial images," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 3791-3798.

[26] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740-755.

[27] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6154-6162.

[28] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980-2988.

[29] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6568-6577.

[30] S. Zhang, C. Chi, Y. Yao, Z. Lei and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9756-9765.

[31] J. Chen, S.-H. Kao, H. He, W. Zhuo, S. Wen and C.-H. Lee, "Run, don't walk: Chasing higher FLOPS for faster neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12021-12031.

[32] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue and B. Zhang, "Rethinking mobile block for efficient attention-based models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1389-1400.

[33] J. Li, Y. Wen and L. He, "SCConv: Spatial and channel reconstruction convolution for feature redundancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 6153-6162.

[34] X. Li, W. Wang, X. Hu, J. Li, J. Tang and J. Yang, "Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11627-11636.

[35] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu and X. Wang, "YOLO-FaceV2: A scale and occlusion aware face detector," *Pattern Recognit.*, vol. 155, Art. no. 110714, 2024.

[36] J. Zhang, Z. Chen, G. Yan, Y. Wang and B. Hu, "Faster and lightweight: An improved YOLOv5 object detector for remote sensing images," *Remote Sens.*, vol. 15, no. 20, Art. no. 4974, 2023.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618-626.