Cil-FPN: A New Paradigm for Efficient Feature Fusion for UAV Object Detection

Xiyu Pan, Kai Xiong, and Jianjun Li*
Central South University of Forestry and Technology, Changsha, China
{20231100392, 20231100410, T20010539}@csuft.edu.cn

***Abstract*—Object detection in UAV imagery faces persistent challenges from small objects, dense distributions, and complex backgrounds. Conventional detectors tend to lose fine-grained spatial cues due to repeated down-sampling, limiting small-object perception. To address this, we introduce the Cross-layer Lightweight Feature Pyramid Network (Cil-FPN), which enhances detection by combining shallow high-resolution features with deep semantic cues. Within Cil-FPN, the Multibranch Competitive Frequency-guided Module (MCFM) integrates global context pooling, frequency-based edge enhancement, and branch competition for efficient and discriminative feature fusion. Designed as a lightweight plug-and-play component, Cil-FPN can be seamlessly incorporated into mainstream detectors such as YOLO11, RT-DETR, and DEIM. Extensive experiments on four challenging benchmarks—`VisDrone2019`, `CODrone`, `HIT-UAV`, and `CARPK`—demonstrate improved accuracy with reduced computational cost. For instance, YOLO11 equipped with Cil-FPN and MCFM improves mAP@0.5 from 31.6 to 33.2 on `VisDrone2019` with only 3.7M parameters. To facilitate reproducibility, the code is provided at https://github.com/panxiyu2001/Cil-FPN.**

***Index Terms*—UAV object detection, Feature fusion, Small object detection, Remote sensing**

I. Introduction

Object detection in UAV imagery is critical for applications like urban monitoring [1] and disaster response [2]. However, as shown in Fig. 1, extreme object density and tiny scales pose persistent hurdles [3]. Traditional detectors relying on conventional Feature Pyramid Networks (FPNs) [4] often prioritize deeper layers for semantic abstraction, which unfortunately suppresses fine-grained spatial cues [5], leading to missed detections of small objects.

To address this, we propose the Cross-layer Lightweight Feature Pyramid (**Cil-FPN**), a plug-and-play component that incorporates low-level backbone features to retain spatial details while integrating higher-level semantics. Within Cil-FPN, we introduce the Multi-branch Competitive Frequency-guided Module (**MCFM**), which leverages frequency-domain edge cues and inter-branch competition for discriminative fusion. By combining global context pooling with high-frequency information, MCFM effectively handles small-object perception in complex UAV scenes.

We integrate our modules into YOLO11 [6], RT-DETR [7], and DEIM [8]. Evaluations on `VisDrone2019`, `CODrone`, `HIT-UAV`, and `CARPK` demonstrate that our method consistently improves accuracy while significantly reducing parameters and GFLOPs (Fig. 2). Our contributions include:

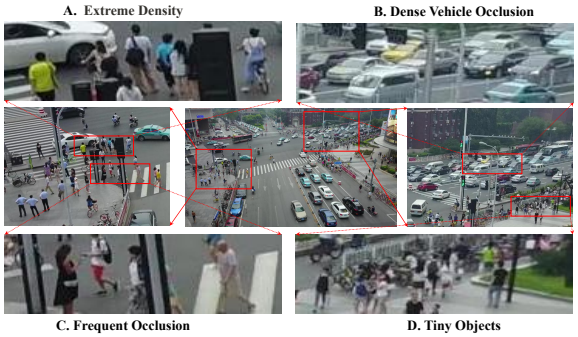

Figure 1. Core challenges in VisDrone2019. Insets illustrate the co-existing issues of (a)-(c) extreme density and severe occlusion across crowds/vehicles, and (d) the prevalence of tiny objects.

- **Cil-FPN Architecture:** A shallow-focused paradigm using a *scale-migration strategy* to reallocate capacity to high-resolution layers, preventing information collapse for tiny objects without the cost of high-resolution detection heads.
- **MCFM Module:** A competition-driven fusion mechanism that integrates frequency-domain structural priors to prioritize informative features over background noise.
- **Extensive Validation:** Our method achieves superior accuracy-efficiency trade-offs across CNN and Transformer-based detectors on four challenging UAV benchmarks.

II. Related Work

UAV Object Detection: Standard detectors [7], [9]–[11] often yield suboptimal results in UAV imagery due to extreme scale variations and tiny objects [3]. Existing specialized solutions like slicing-aided inference [12] or super-resolution [13] improve perception but introduce heavy computational overhead or latency. Unlike these, our Cil-FPN focuses on architectural efficiency to retain small-object cues without external image processing.

Feature Pyramid Networks: Multi-scale fusion is standard for multi-size object detection. While FPN variants [4], [14]–[17] have refined information flow, they largely rely on deep, low-resolution layers (e.g., P5). For UAV scenarios, objects as small as 30×30 pixels collapse in P5, causing irreversible information loss [5]. Cil-FPN addresses this by discarding

*Corresponding author.

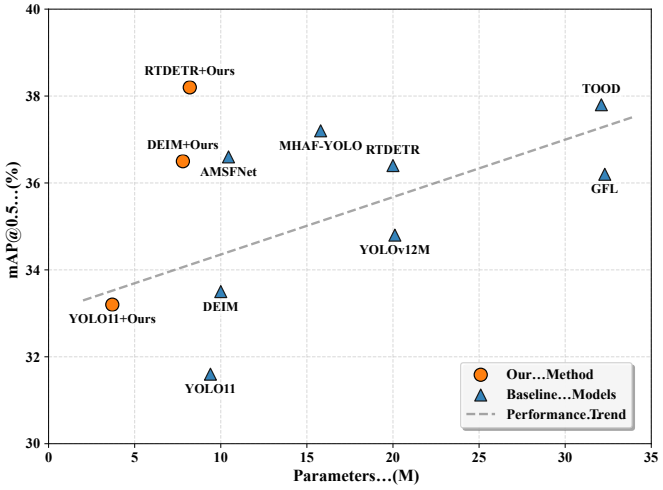

Figure 2. Accuracy-parameter trade-off on `VisDrone`. Our method achieves superior performance with a smaller model size.

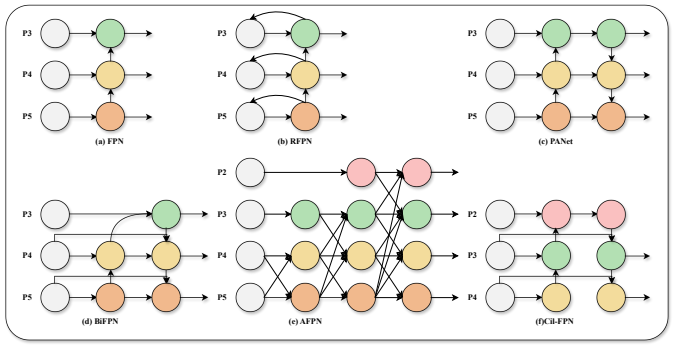

Figure 3. Evolution of Feature Pyramid Architectures. (a-d) Traditional designs (FPN, RFPN, PANet, BiFPN) predominantly operate on low-resolution P3–P5 scales, causing tiny objects to vanish in the deep P5 layer (stride 32). (e) AFPN attempts to include high-resolution P2 features but suffers from excessive computational complexity due to dense connectivity. (f) Our proposed **Cil-FPN** introduces a distinct *"Shallow-Focus"* paradigm: it discards the coarse P5 layer to prune redundancy and structurally integrates the high-resolution P2 layer. This design efficiently resolves the small object information loss problem without the heavy overhead of dense fusion.

coarse P5 features and shifting the focus to high-resolution shallow layers (P2) to preserve fine-grained details.

Frequency-domain Learning: Frequency domain cues, such as DCT or Laplacian operators, offer unique perspectives for capturing edges and textures [18]–[20]. Despite its potential for structural enhancement, frequency guidance is rarely exploited in lightweight detection. Our MCFM bridges this gap by explicitly using high-frequency edge maps to guide attention, providing a principled way to handle dense small-object distributions.

III. Methodology

A. Motivations

Feature fusion is vital for small, dense objects in UAV imagery [21]. Conventional designs like PANet [14] and BiFPN [15] primarily exploit deep backbone features, which often suppress fine-grained spatial details. To address this, Cil-FPN (Fig. 3) incorporates shallower feature maps to retain critical edge and texture cues while maintaining semantic abstraction through intermediate layers [22]. Furthermore, Cil-FPN integrates the MCFM to enhance the discriminability of fused representations by combining context and frequency-domain information in a lightweight manner. Detailed technical descriptions follow.

B. Technical Details

Cil-FPN Architecture. As visually compared in Fig. 3, the design philosophy of feature pyramids significantly determines detection performance. Traditional architectures like FPN, RFPN, PANet, and BiFPN (Fig. 3a-d) rely heavily on deep semantic layers (P3–P5). While P5 provides rich semantics for large objects in natural scenes (e.g., COCO), it is detrimental to UAV imagery: a 32×32 pixel object collapses to a single feature point in P5 (stride 32), leading to irreversible *feature collapse*. Although recent advanced designs like AFPN (Fig. 3e) attempt to incorporate shallower features (P2), they resort to complex dense connections, resulting in high latency and computational burden unsuited for edge deployment.

To address these limitations, Cil-FPN proposes a novel "Scale-Migration" strategy. Instead of simply adding layers which increases cost, we perform a strategic trade-off: we discard the coarse P5 layer—which contributes little to tiny object detection—and reallocate computational resources to the high-resolution P2 layer (stride 4). This results in a focused P2–P4 pyramidal system (Fig. 3f). By centering the architecture on shallow layers, Cil-FPN retains critical high-frequency spatial details (e.g., edges and textures of tiny cars or crowds) that are typically washed out in traditional P3-P5 baselines.

Structurally, Cil-FPN draws inspiration from the efficient cross-scale connections of BiFPN but tailors the topology specifically for UAV constraints. First, to maximize feature preservation, we incorporate lateral skip connections (indicated by the horizontal arrows jumping to the output nodes in Fig. 3f). These connections directly transmit original backbone features to the final fusion stage, preventing the dilution of fine-grained cues during repeated sampling. Second, unlike standard pyramids that output predictions at all levels, we devise a "Feature-Injection, Selective-Prediction"strategy. While the computational-heavy high-resolution P2 features participate extensively in the top-down and bottom-up fusion to inject spatial details into P3 and P4, they are *excluded* from the final detection head. The model ultimately outputs predictions only from P3 and P4. This design successfully harvests the spatial precision of P2 to refine the semantic layers without incurring the massive computational cost of performing dense predictions on the high-resolution P2 grid. Finally, all fusion nodes employ our MCFM to dynamically weight these multi-source inputs, ensuring distinct target features are prioritized over background noise.

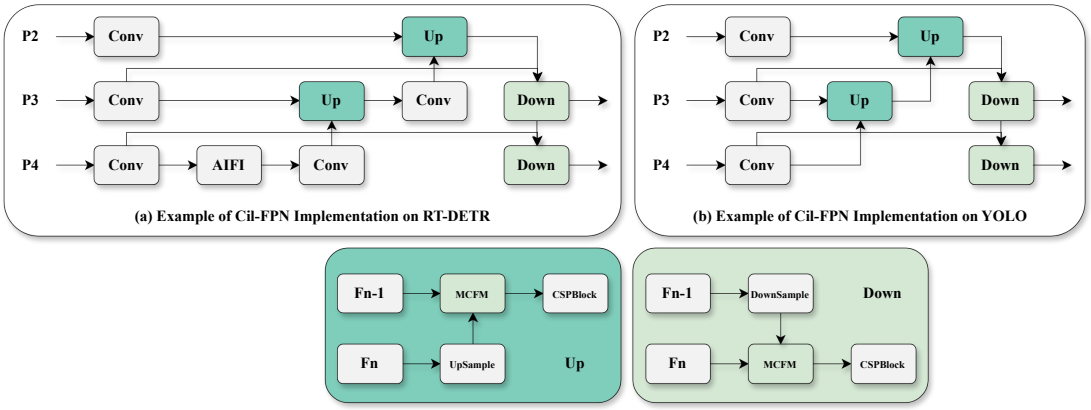

Figure 4. Implementation details of Cil-FPN on different detectors. (a) Integration with **RT-DETR**, preserving the AIFI module for P4. (b) Integration with **YOLO**, replacing the standard neck. The bottom row details the specific "Up" and "Down" blocks, illustrating how MCFM serves as the core fusion unit to aggregate multi-scale features effectively.

Implementation on Diverse Detectors. Fig. 4 demonstrates the seamless integration of Cil-FPN into both CNN-based (YOLO) and Transformer-based (RT-DETR) architectures.

- **For RT-DETR (Fig. 4a):** We retain the Adaptive Interaction Fusion Interface (AIFI) at the highest semantic level (P4) to capture long-range dependencies using intra-scale self-attention. The subsequent cross-scale fusion is handled by Cil-FPN, where the "Up" and "Down" blocks leverage MCFM to merge encoder features with the re-parameterized backbone features.
- **For YOLO (Fig. 4b):** Cil-FPN replaces the standard PANet neck. The backbone features (P2, P3, P4) are first aligned via 1×1 convolutions. The fusion process employs a dual-path strategy: the *Top-down* path enhances semantic consistency, while the *Bottom-up* path recovers edge information.

In both implementations, the distinct "Up" module combines up-sampling with MCFM-based fusion, while the "Down" module utilizes down-sampling followed by MCFM, ensuring that multi-scale interactions are both lightweight and discriminative.

MCFM. Although Cil-FPN introduces shallow features to preserve fine-grained spatial details, simply aggregating multi-scale features via summation or concatenation treats all inputs equally, ignoring their distinct semantic and structural importance.

Unlike spatial-domain attention mechanisms that primarily respond to semantic saliency, high-frequency components encode structural cues such as object boundaries and texture discontinuities, which are particularly informative for distinguishing densely packed small objects in UAV imagery. In such scenarios, semantic responses alone often become ambiguous due to severe occlusion and scale compression.

Motivated by this observation, we introduce explicit frequency guidance into feature fusion. Rather than learning complex spectral transformations, we adopt a fixed Laplacian operator as a computationally efficient approximation of high-pass filtering. This choice provides stable edge-sensitive responses without introducing additional learnable parameters, making it well suited for lightweight detection frameworks.

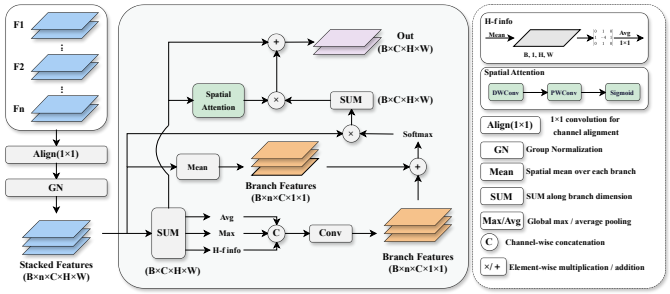

Figure 5. The architecture of the proposed Multi-branch Competitive Frequency-guided Module (MCFM). The module aligns multi-scale inputs and employs a competitive mechanism driven by global pooling and high-frequency edge cues (H-f info) to dynamically reweight features. A spatial attention branch further refines the fused representation.

Given multi-branch feature maps $\{F_i \in \mathbb{R}^{C_{in} \times H \times W}\}_{i=1}^{n}$, we first align their channel dimensions via a 1×1 convolution and standardize them using Group Normalization (GN). These aligned features are stacked to form a unified representation. To capture comprehensive contextual information, we compute a shared consensus feature $F_{\text{sum}} = \sum_{i=1}^{n} F_i$.

1) Frequency-guided Context Encoding: Unlike standard attention mechanisms that only use Max/Avg pooling, MCFM explicitly incorporates edge information to enhance small object localization. As shown in the "H-f info" block of Fig. 5, we apply a Laplacian operator with a fixed kernel $K_{lap} = [[0, 1, 0], [1, -4, 1], [0, 1, 0]]$ to F_{sum}, followed by global average pooling. This high-frequency descriptor is concatenated with the global average and max-pooled features of F_{sum} to generate a shared context vector V_{shared}.

2) Competitive Branch Selection: To dynamically adjudicate the importance of each branch, we formulate a competitive selection mechanism. First, we generate branch-specific descriptors V_{local}^{i} via Global Average Pooling (GAP) on each aligned feature F_i. To integrate global guidance with local specificity,

Table I
PERFORMANCE COMPARISON OF VARIOUS MODELS.

Model	VisDrone mAP@0.5	CODrone mAP@0.5	Params(M)	GFLOPs
TOOD [23]	37.8	32.1	32.1	199
GFL [24]	36.2	30.1	32.3	206
AMSFNet [25]	36.6	29.8	10.45	39.8
YOLOv12M [26]	34.8	29.6	20.11	67.2
MHAF-YOLO [27]	37.2	32.1	15.8	67.6
YOLO11 [6]	31.6	26.9	9.4	21.5
DEIM [8]	33.5	26.5	10	25.0
RTDETR [7]	36.4	31.9	20	60.0
YOLO11+Ours	33.2	28.7	3.7	17.7
DEIM+Ours	36.5	30.5	7.8	22
RTDETR+Ours	38.2	33.9	8.2	53.1

we model the interaction between the shared frequency-aware context V_{shared} and the local descriptor via element-wise addition. The final attention logits are then projected using a 1×1 convolution. To strictly control the sharpness of the branch distribution, we employ a learnable temperature-scaled Softmax, formulated as:

$$\alpha^i = \text{Softmax}\left(\frac{\text{Conv}(V_{shared} + V_{local}^i)}{\tau}\right) \tag{1}$$

where $\tau = 0.5 + \text{Softplus}(\tau_{\text{learn}})$. Here, τ acts as a regulating factor: a smaller τ approximates a "winner-takes-all" selection (sharpening the weights towards the most informative branch), while a larger τ yields a smoother, uniform distribution. This learnable design allows the network to adaptively switch between selective focus and feature aggregation during training.

3) Spatial Refinement and Fusion: While α^i reweights features in a channel-wise manner, spatial discriminability is equally crucial for dense scenes. To this end, we generate a spatial attention map M_s from F_{sum} using a Depthwise–Pointwise Convolution (DW–PW) pair followed by a Sigmoid function. The final output F_{out} integrates competitively weighted branches, spatial modulation, and a residual connection:

$$F_{out} = \left(\sum_{i=1}^{n} \alpha^i \cdot F_i\right) \odot M_s + F_{\text{sum}}. \tag{2}$$

This formulation effectively suppresses background noise while preserving fine-grained details of small targets.

In essence, MCFM addresses a fundamental question in multi-scale feature fusion: given multiple cross-scale features, which one should dominate under dense and cluttered scenes? Rather than uniformly aggregating all inputs, MCFM enforces explicit competition among branches, allowing the most informative representation to prevail. Meanwhile, frequency-guided cues provide an explicit structural prior for small-object boundaries, steering the competition toward edge-sensitive and detail-preserving features.

IV. EXPERIMENT ANALYSIS

Experimental Setup: All models were trained from scratch on an NVIDIA RTX 4060Ti (PyTorch 2.2) to isolate architectural contributions. We set 400 epochs for CNN-based models (YOLO11) and 200 for Transformer-based models (RT-DETR, DEIM). We evaluate on four UAV benchmarks: `VisDrone2019` [28] (urban occlusions), `CODrone` [29] (scale variations), `HIT-UAV` [30] (tiny targets), and `CARPK` [31] (dense counting).

Table II
GENERALIZATION PERFORMANCE ON FOUR DATASETS. **MAP@0.5** AND **MAP@0.5:0.95** DENOTE DETECTION ACCURACY. OUR METHOD MAINTAINS HIGHER INFERENCE SPEED (FPS) WHILE CONSISTENTLY IMPROVING ACCURACY. GAINS ARE MARKED IN RED.

Dataset	Method	mAP@0.5	mAP @0.5:0.95	FPS
VisDrone	YOLO11	31.6	18.1	95.0
	+ **Ours**	**33.2** (+1.6)	**19.3** (+1.2)	**112.5**
	RT-DETR	36.4	20.6	74.0
	+ **Ours**	**38.2** (+1.8)	**22.3** (+1.7)	**86.2**
	DEIM	33.5	19.5	88.0
	+ **Ours**	**36.5** (+3.0)	**20.9** (+1.4)	**101.8**
CoDrone	YOLO11	26.9	13.6	95.0
	+ **Ours**	**28.7** (+1.8)	**15.1** (+1.5)	**112.5**
	RT-DETR	31.9	16.2	74.0
	+ **Ours**	**33.9** (+2.0)	**17.8** (+1.6)	**86.2**
	DEIM	26.5	14.1	88.0
	+ **Ours**	**30.5** (+4.0)	**16.4** (+2.3)	**101.8**
HIT-UAV	YOLO11	80.7	45.2	95.0
	+ **Ours**	**82.2** (+1.5)	**46.5** (+1.3)	**112.5**
	RT-DETR	82.1	47.6	74.0
	+ **Ours**	**84.0** (+1.9)	**49.2** (+1.6)	**86.2**
	DEIM	81.9	46.8	88.0
	+ **Ours**	**83.1** (+1.2)	**48.4** (+1.6)	**101.8**
CARPK	YOLO11	81.6	55.4	95.0
	+ **Ours**	**83.7** (+2.1)	**57.2** (+1.8)	**112.5**
	RT-DETR	84.1	58.2	74.0
	+ **Ours**	**86.5** (+2.4)	**60.4** (+2.2)	**86.2**
	DEIM	83.6	57.5	88.0
	+ **Ours**	**86.1** (+2.5)	**59.6** (+2.1)	**101.8**

A. Comparison with State-of-the-Art Methods

As shown in Tables I and II, our modules consistently enhance strong baselines while reducing overhead. On `VisDrone2019`, **YOLO11+Ours** achieves 33.2 mAP with only **3.7M** parameters and **17.7** GFLOPs—an 88% reduction in size and 91% in complexity compared to TOOD [23], with comparable accuracy. Results on `HIT-UAV` and `CARPK` confirm our method's robustness in high-altitude and dense scenarios. The dramatic reduction in GFLOPs suggests CiI-FPN is ideal for edge devices like NVIDIA Jetson, where the removal of the P5 layer mitigates critical memory bottlenecks.

B. Ablation Study

Quantitative results on `VisDrone2019` and `CODrone` (Table III) verify our contributions:

Table III

ABLATION STUDIES. MAP@0.5 AND MAP@0.5:0.95 ARE THE DETECTION ACCURACY AT DIFFERENT IOU THRESHOLDS. 'PARAMS' (M) IS THE NUMBER OF MODEL PARAMETERS. 'GFLOPS' MEANS A MODEL'S COMPUTATIONAL COMPLEXITY, REPRESENTING THE BILLIONS OF FLOATING-POINT OPERATIONS NEEDED FOR ONE FORWARD PASS. '✓' AND '✗' DENOTE WHETHER THE METHODS INCLUDE THE CIL-FPN AND MCFM.

Model	Cil-FPN	MCFM	VisDrone2019		CODrone		Params (M)	GFLOPs
			mAP@0.5	mAP @0.5:0.95	mAP@0.5	mAP @0.5:0.95		
YOLO11	✗	✗	31.6	18.1	26.9	13.6	9.4	21.5
	✓	✗	32.8	18.9	28.4	15.0	**3.0**	18.1
	✗	✓	33.0	19.1	28.7	14.9	10.8	21.7
	✓	✓	**33.2**	**19.3**	**28.7**	**15.1**	3.7	**17.7**
RT-DETR	✗	✗	36.4	20.6	31.9	16.2	20.0	60.0
	✓	✗	38.0	22.1	33.2	17.4	**6.1**	48.4
	✗	✓	37.8	21.9	**34.1**	**18.2**	21.6	57.1
	✓	✓	**38.2**	**22.3**	33.9	17.8	8.2	53.1
DEIM	✗	✗	33.5	19.5	26.3	14.1	10.0	25.0
	✓	✗	35.3	20.1	28.6	15.4	**7.6**	23.0
	✗	✓	35.6	20.3	28.9	15.9	11.5	22.3
	✓	✓	**36.5**	**20.9**	**30.5**	**16.4**	7.8	**22.0**

Table IV

SENSITIVITY ANALYSIS OF THE INITIAL TEMPERATURE PARAMETER (τ_{init}) ON THE VISDRONE2019 VALIDATION SET ACROSS THREE ARCHITECTURES. THE DEFAULT SETTING ($\tau = 0.7$) CONSISTENTLY YIELDS THE BEST TRADE-OFF FOR ALL MODELS.

Initial Temp. (τ)	YOLO11		RT-DETR		DEIM	
	mAP50	mAP	mAP50	mAP	mAP50	mAP
0.5	33.0	19.1	38.0	22.1	36.3	20.7
0.7 (Ours)	**33.2**	**19.3**	**38.2**	**22.3**	**36.5**	**20.9**
1.0	33.1	19.2	38.1	22.2	36.4	20.8
1.5	32.8	19.0	37.8	22.0	36.1	20.5

Effectiveness of Cil-FPN: Replacing the standard neck with Cil-FPN reduces YOLO11 parameters by 68% and GFLOPs by 16% while improving mAP by 1.2%. This proves that focusing on P2 layers is more effective than deep P5 layers for UAV imagery.

Impact of MCFM: MCFM boosts YOLO11 mAP to 33.0 by refining features via frequency cues. While it slightly increases parameters if used alone, its synergy with Cil-FPN provides the optimal accuracy-efficiency trade-off.

Synergy and Generalization: The full method slashes RT-DETR parameters by 59% (+1.8 mAP) and delivers a 3.0 mAP gain for DEIM. This confirms the modules' effectiveness across both CNN and Transformer architectures.

Sensitivity of τ: Analysis of the temperature parameter τ (Eq. 1) identifies $\tau_{init} = 0.7$ as the optimal setting across all baselines (Table IV), balancing branch competition with gradient stability.

C. Qualitative Visualization Analysis

Detection Robustness: Fig. 6 illustrates the comparative detection results on the VisDrone dataset, covering complex scenes such as dense parking lots, intersections, and low-light environments. We employ a differential color scheme: green for consensus detections, red for objects detected exclusively by our method, and blue for those exclusive to the baseline. As observed, our model successfully retrieves a substantial

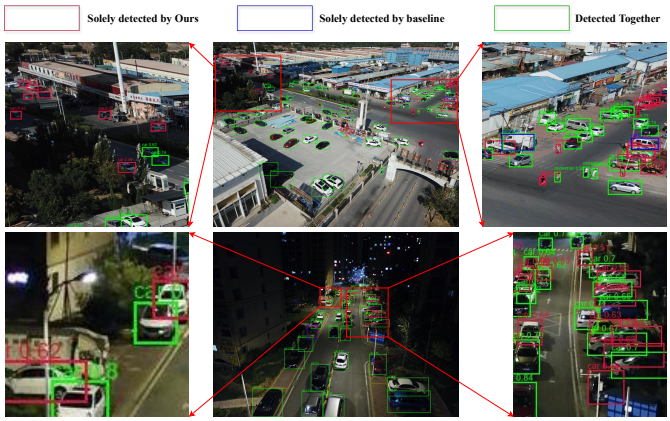

Figure 6. Visual Comparison of the Original Model and the Improved Model Using Our Cil-FPN Method on the VisDrone Dataset.

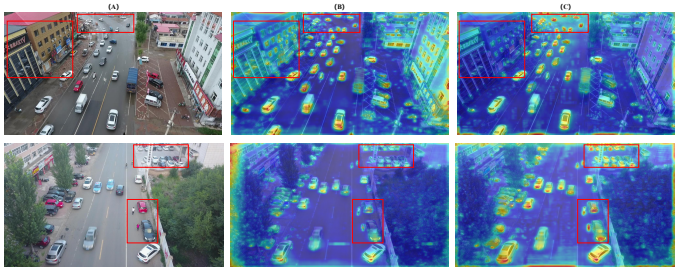

Figure 7. Grad-CAM++ attention maps. (A) Input (B) Baseline Models (C) Ours Method

number of small and densely packed targets (highlighted in red) that are missed by the baseline, particularly in the crowded vehicle clusters and distant views. The scarcity of blue boxes further indicates that our method retains the vast majority of valid detections found by the baseline while significantly improving recall for hard-to-detect objects. This confirms that the shallow feature preservation in Cil-FPN effectively prevents small objects from being lost during down-sampling.

Attention Mechanism: To further investigate the internal

decision-making process, we visualize the attention maps using Grad-CAM++ in Fig. 7. Comparing the baseline (Column B) with our method (Column C), it is evident that the baseline often exhibits diffuse attention, struggling to separate small targets from the background (e.g., the distant cars in the red boxes). In contrast, our method generates highly concentrated and intense activation regions (Column C) that precisely align with the objects of interest. This enhanced focus validates the contribution of the MCFM module, which effectively utilizes frequency-guided cues and branch competition to suppress background noise and enhance discriminability for tiny targets.

V. Conclusion

This work presents a lightweight yet powerful paradigm for UAV object detection, addressing the persistent challenges of scale variation and occlusion. We propose Cil-FPN to reinvigorate fine-grained spatial cues from shallow layers, together with MCFM, a frequency-guided module that leverages competitive learning to refine feature representations. Unlike conventional approaches that rely predominantly on deep semantic features, our method achieves a more balanced trade-off between spatial precision and semantic abstraction. Extensive experiments on four challenging UAV benchmarks demonstrate that the proposed plug-and-play modules consistently improve the accuracy of state-of-the-art detectors (e.g., YOLO11 , RT-DETR and DEIM) while substantially reducing model parameters and computational cost.

It is worth noting that the removal of the coarse P5 layer introduces a deliberate inductive bias toward small- and medium-scale objects, which dominate UAV imagery. Although extremely large objects in natural-scene datasets may benefit from deeper semantic representations, our empirical results across four UAV benchmarks suggest that such cases are relatively rare and do not outweigh the benefits of preserving high-resolution spatial information. Future work will explore adaptive scale-activation strategies that dynamically reintroduce coarse features when necessary, further improving the flexibility of the proposed framework.

References

[1] J. Wang *et al.*, "Tiny object detection in aerial images," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 3791–3798.

[2] H. Ijaz *et al.*, "A uav-assisted edge framework for real-time disaster management," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[3] X. Zhang *et al.*, "Remote sensing object detection meets deep learning: A metareview of challenges and advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 4, pp. 8–44, 2023.

[4] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 936–944.

[5] Y. Ma *et al.*, "Seemore: a spatiotemporal predictive model with bidirectional distillation and level-specific meta-adaptation," *Science China Information Sciences*, vol. 67, no. 1, p. 182104, 2024. [Online]. Available: https://doi.org/10.1007/s11432-022-3859-8

[6] G. Jocher *et al.*, "Yolo11 by ultralytics," GitHub, 2025, https://github.com/ultralytics/ultralytics.

[7] Y. Zhao and others., "Detrs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 16 965–16 974.

[8] S. Huang *et al.*, "Deim: Detr with improved matching for fast convergence," 2025. [Online]. Available: https://arxiv.org/abs/2412.04234

[9] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[10] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[11] N. Carion *et al.*, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[12] F. C. Akyon *et al.*, "Slicing aided hyper inference and fine-tuning for small object detection," in *2022 IEEE international conference on image processing (ICIP)*. IEEE, 2022, pp. 966–970.

[13] M. Nikouei *et al.*, "Small object detection: A comprehensive survey on challenges, techniques and real-world applications," *arXiv preprint arXiv:2503.20516*, 2025.

[14] S. Liu *et al.*, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 8759–8768.

[15] M. Tan *et al.*, "Efficientdet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 10 778–10 787.

[16] G. Yang *et al.*, "Afpn: Asymptotic feature pyramid network for object detection," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2023, pp. 2184–2189.

[17] S. Qiao *et al.*, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 213–10 224.

[18] Z. Qin *et al.*, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[19] X. Pan *et al.*, "Adaptive edge-aware detection with lightweight multi-scale fusion," *Electronics*, vol. 15, no. 2, 2026. [Online]. Available: https://www.mdpi.com/2079-9292/15/2/449

[20] K. Xu *et al.*, "Learning in the frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1740–1749.

[21] Y. Xiao *et al.*, "A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[22] Z. Hu *et al.*, "Dome-detr: Detr with density-oriented feature-query manipulation for efficient tiny object detection," 2025.

[23] C. Feng *et al.*, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3490–3499.

[24] X. Li *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," 2020. [Online]. Available: https://arxiv.org/abs/2006.04388

[25] Y. Chen *et al.*, "Axis-squeeze and multirouting scale-adaptive fusion network for remote sensing images object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.

[26] Y. Tian *et al.*, "Yolov12: Attention-centric real-time object detectors," 2025. [Online]. Available: https://arxiv.org/abs/2502.12524

[27] Z. Yang *et al.*, "Mhaf-yolo: Multi-branch heterogeneous auxiliary fusion yolo for accurate object detection," 2025. [Online]. Available: https://arxiv.org/abs/2502.04656

[28] D. Du and others., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South), 2019, pp. 213–226.

[29] K. Ye *et al.*, "More clear, more flexible, more precise: A comprehensive oriented object detection benchmark for uav," 2025. [Online]. Available: https://arxiv.org/abs/2504.20032

[30] J. Suo *et al.*, "Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," *Scientific Data*, vol. 10, p. 227, 2023.

[31] M.-R. Hsieh *et al.*, "Drone-based object counting by spatially regularized regional proposal network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4145–4153.