FR-YOLO: A Focus-and-Reconstruct Mechanism for Lightweight Small Object Detection in Drone Imagery

Hongwei Liu[1], Rui Zhu[1], Shisheng Tang[1], Liwei Sun[1], Haohan Ding[1,2,*]

[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
[2]Science Center for Future Foods, Jiangnan University, Wuxi, China
13851061944@163.com

Abstract. Automatic detection of small objects in UAV imagery is a challenging problem that is of great interest in aerial surveillance and intelligent transport. The detection model must cope well with feature degradation of tiny targets, drastic scale variations, and strict constraints on onboard computational resources. This paper proposes a lightweight attention-based network, called FR-YOLO, to address the "focus" and "reconstruct" challenges in small object detection. We introduce two novel components: the Local Feature Enhancement (LFE) module to precisely suppress background noise via spatial attention, and the Content-aware Feature Reassembly (CFR) module to rectify spatial feature misalignment and recover fine-grained details during upsampling. The proposed modules are seamlessly integrated into the YOLO11n backbone. Extensive experiments on the authoritative VisDrone2019 benchmark indicate that FR-YOLO outperforms the YOLO11n baseline by a significant margin of 1.5% and surpasses the widely adopted YOLOv8n by 0.7% (achieving 36.4% mAP), all while maintaining a 16% smaller model size (2.53M parameters) compared to YOLOv8n. We have made the code available to the public at: https://github.com/Liu999hongwei/FR-YOLO.

Keywords: UAV object detection, spatial attention, feature alignment, YOLO, lightweight network.

1 Introduction

Automatic detection of ground targets constitutes the foundational perception layer of the low-altitude economy. Owing to their high mobility, Unmanned Aerial Vehicles (UAVs) have become indispensable for critical missions [1]. However, traditional surveillance frameworks suffer from severe latency and bandwidth bottlenecks [2], while drastic viewpoint variations in UAV imagery pose significant challenges to standard algorithms [3].

Despite remarkable progress in deep learning [4] and various one-stage and two-stage detectors [5,6], complex UAV detection models exceed the power and memory constraints of onboard edge platforms. Consequently, highly efficient lightweight detectors (e.g., the YOLO series) are widely adopted. Nevertheless, their aggressive downsampling strategies (e.g., a 32-fold stride) cause tiny objects—often smaller than

32×32 pixels—to rapidly attenuate in deep feature maps, resulting in an irreversible loss of critical geometric information.

This feature loss severely exacerbates the "accuracy-efficiency" dilemma in UAV-based small object detection [7], where extreme environmental interference demands robust feature extraction [8] against strictly lightweight architectural constraints [9]. Specifically, generic lightweight detectors struggle with two critical bottlenecks during deployment (Fig. 1): First (Fig. 1a), high-frequency background noise from complex environments (e.g., illumination changes and tree shadows) easily overwhelms the inherently weak signals of tiny targets, causing missed detections. Second (Fig. 1b), in severely occluded crowds, standard content-agnostic nearest-neighbor upsampling interpolates solely based on spatial coordinates. This completely disregards semantic context, inducing fatal semantic misalignment and target merging, thereby severely degrading localization accuracy [10]. To systematically address these challenges, we propose a "Focus-and-Reconstruct" mechanism precisely tailored for UAV tiny object detection.

Fig. 1. Motivation of FR-YOLO. Generic detectors suffer from: (a) noise interference masking tiny objects, and (b) semantic misalignment merging crowded targets. Our method resolves both.

To systematically address the challenges of UAV-based tiny object detection, our principal contributions are three-fold:

(1) We propose FR-YOLO, a lightweight framework featuring a novel "Focus-and-Reconstruct" mechanism. It effectively mitigates feature degradation for tiny objects while maintaining real-time inference on edge devices.
(2) We design two synergistic, plug-and-play modules: Local Feature Enhancement (LFE) utilizes spatial attention to filter background noise (Focus), while Content-aware Feature Reassembly (CFR) employs dynamic kernels to rectify spatial misalignment and restore fine-grained details (Reconstruct).
(3) Extensive evaluations on the VisDrone2019 benchmark demonstrate that FR-YOLO achieves an optimal accuracy-speed trade-off, significantly outperforming state-of-the-art baselines. Our source code is publicly released.

2 Related Work

2.1 Object Detection in UAV Imagery

Generic object detection algorithms have evolved from two-stage paradigms (e.g., Faster R-CNN [5]) to one-stage paradigms (e.g., SSD and the YOLO series [11]). In particular, the YOLO framework has become the de facto standard in industry due to its superior balance between inference speed and detection accuracy. From YOLOv5 to the latest YOLO11, the network architecture has been continuously optimized through the introduction of efficient aggregation modules (e.g., C2f, C3k2) and decoupled heads [12]. However, directly transferring these generic detectors to UAV scenarios presents formidable challenges. Unlike natural scene datasets such as COCO, UAV datasets like VisDrone are characterized by extremely tiny object scales, highly dense distributions, and complex background interference [13]. Although YOLO11 performs exceptionally well on normal-sized objects, its aggressive downsampling strategy—adopted in pursuit of extreme lightweightness—causes the fine-grained features of tiny objects to rapidly attenuate or even completely vanish in deep network layers [14,15]. Therefore, there is an urgent need to develop a lightweight detection architecture tailored specifically to the unique characteristics of UAV scenarios, in order to recover the feature representation of tiny objects under stringent computational constraints.

2.2 Visual Attention Mechanisms

Attention mechanisms have been widely proven effective in various computer vision tasks for enhancing feature representation through the adaptive recalibration of feature weights. Classical methods (e.g., CBAM [16]) utilize sequential channel and spatial attention mechanisms to highlight salient regions while maintaining relatively low parameter overhead. However, most lightweight detectors (including the YOLO series), in their pursuit of minimalist network structures and extreme inference speeds, typically compromise on or completely discard such explicit spatial attention modules. This design trade-off renders them highly vulnerable to the high-frequency background noise prevalent in aerial imagery (e.g., complex tree textures, building shadows), making them highly susceptible to false positives [17]. To remedy this deficiency, this paper designs the Local Feature Enhancement (LFE) module based on the spatial attention paradigm. By strategically embedding it into the backbone network, our model can explicitly suppress interference from non-target regions at an early stage, thereby achieving precise "Focus" on tiny objects with minimal additional parameters.

2.3 Feature Upsampling and Multi-scale Fusion

In the Feature Pyramid Network (FPN) of object detection architectures, feature upsampling is a crucial operation for achieving effective cross-scale information fusion. The traditional nearest-neighbor interpolation adopted by default in the YOLO series is fundamentally a content-agnostic strategy. It relies purely on the spatial positions of

pixels for feature expansion, which inevitably leads to severe semantic misalignment and blurred object boundaries when processing dense, tiny UAV targets [18]. In recent years, advanced operators such as CARAFE [19] have significantly improved upsampling quality by predicting spatially varying, content-aware convolutional kernels to guide feature reassembly. However, such dynamic reassembly strategies are typically accompanied by high computational complexity and memory overhead, and thus have not been widely applied in resource-constrained real-time UAV detection [20]. Inspired by this limitation, this paper proposes the novel CFR module. This module is customized for lightweight adaptation to UAV scenarios; it utilizes the content-aware paradigm to dynamically rectify feature misalignment and precisely recover the fine-grained geometric details of tiny objects lost during downsampling, thereby successfully resolving the "Reconstruct" dilemma for small targets at an extremely low computational cost.

3 Method

This section systematically elaborates on the proposed FR-YOLO framework, specifically engineered to break through the "Focus-and-Reconstruct" bottleneck of tiny object detection in UAV imagery. We first formally define the overall network topology and optimization objective (Section 3.1). Subsequently, we rigorously derive the mathematical mechanisms of the Local Feature Enhancement (LFE) module (Section 3.2) and the Content-aware Feature Reassembly (CFR) module (Section 3.3) from the perspective of low-level tensor operations. The holistic architecture of FR-YOLO is illustrated in Fig. 2.

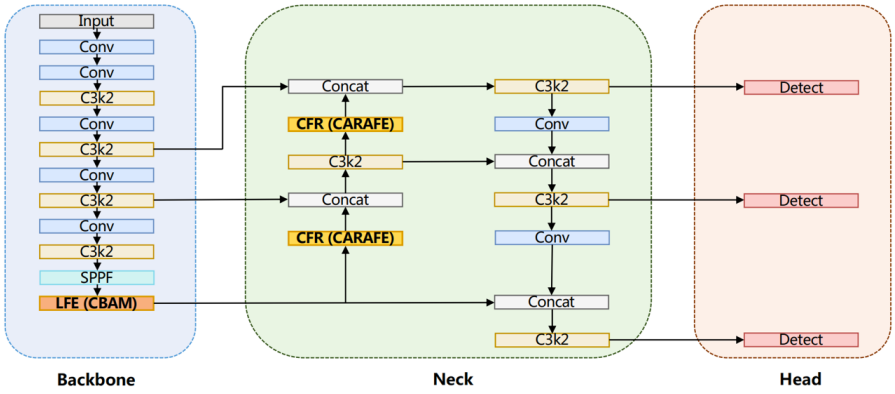

Fig. 2. The overall architecture of the proposed FR-YOLO. It consists of three hierarchical components: the Backbone (blue region), the Neck (green region), and the Head (orange region). The proposed Local Feature Enhancement (LFE) module (highlighted in yellow) is embedded at the deep stage of the backbone to filter high-frequency noise, while the Content-aware Feature Reassembly (CFR) module (highlighted in yellow) is employed in the neck to perform semantic-preserving upsampling. C3k2 and SPPF denote the CSP bottleneck block and the Spatial Pyramid Pooling Fast module, respectively.

3.1 Network Architecture and Optimization Objective

To maximize the feature saliency of tiny targets under the strict constraints of onboard computing resources, we construct the FR-YOLO architecture based on the lightweight YOLO11n. Given an input aerial image $I \in \mathbb{R}^{C\times H\times W}$, the backbone network conducts hierarchical feature extraction via a cascade of C3k2 operators [21]. The extraction process of the multi-scale feature pyramid $\mathcal{P}$ can be formalized as:

$$M_s(F') = \sigma\left(f^{7\times 7}\left([Avg(F'); Max(F')]\right)\right) \tag{1}$$

where $P_i \in \mathbb{R}^{C_i\times\frac{H_0}{2^i}\times\frac{W_0}{2^i}}$ denotes the deep feature map at the downsampling stride of 2^i, and $\mathcal{F}_{backbone}^{(i)}$ represents the non-linear residual mapping at the i-th stage.

To circumvent the irreversible loss of high-frequency signals for tiny objects at the P_5 layer, we embed the LFE module (Φ_{LFE}) at the terminus of the backbone to filter background noise. In the neck stage, to resolve the semantic misalignment during top-down feature propagation, we introduce the CFR module (Φ_{CFR}) to reconstruct high-resolution features. The ultimate optimization objective of the network is driven by a multi-task decoupled loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \lambda_{cls}\sum_{j=1}^{N}\mathcal{L}_{BCE}(p_j,\hat{p}_j) + \lambda_{box}\sum_{j=1}^{N_{pos}}\mathcal{L}_{CIoU}(b_j,\hat{b}_j) + \lambda_{dfl}\sum_{j=1}^{N_{pos}}\mathcal{L}_{DFL}(d_j,\hat{d}_j) \tag{2}$$

where N is the total number of samples, and N_{pos} denotes the positive samples. $\mathcal{L}_{BCE}$, $\mathcal{L}_{CIoU}$, and $\mathcal{L}_{DFL}$, measure the errors for classification, bounding box regression, and Distribution Focal Loss, respectively.λ terms are hyperparameters to balance the losses.

3.2 Local Feature Enhancement (LFE) Module

UAV imagery is frequently saturated with complex background high-frequency noise. The LFE module adaptively recalibrates tensor responses through an explicit spatial-frequency filtering sequence.

Given the deep input feature $F \in \mathbb{R}^{C\times H\times W}$, we first execute global average pooling and max pooling along the spatial dimensions to extract channel descriptors $\mathcal{Z}_{avg}^{c}$, $\mathcal{Z}_{max}^{c} \in \mathbb{R}^{C\times 1\times 1}$. The computation for the c-th channel is precisely formulated as:

$$z_{avg}^{c}(c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F(c,i,j), \quad z_{max}^{c}(c) = \max_{i,j}F(c,i,j) \tag{3}$$

Subsequently, these two descriptors are fed into a shared Multi-Layer Perceptron (MLP). Let $W_0 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ be the weight matrices (where r is the reduction ratio), the generation of the 1D channel attention excitation $M_c \in \mathbb{R}^{C\times 1\times 1}$ is expressed as:

$$M_c(F) = \sigma\left(W_1\delta(W_0 z_{avg}^{c}) + W_1\delta(W_0 z_{max}^{c})\right) \tag{4}$$

where σ is the Sigmoid function and δ represents the ReLU activation. The intermediate feature is updated as $F' = M_c\ (F)\odot F$.

To further focus saliency across the spatial dimension, we compress F' along the channel axis to obtain 2D spatial descriptors $\mathcal{Z}_{avg}^{s}$, $\mathcal{Z}_{max}^{s} \in \mathbb{R}^{1\times H\times W}$. The 2D spatial

attention map $M_s \in \mathbb{R}^{1\times H\times W}$ is mapped via a convolution operation $f^{7\times7}$ with a 7×7 kernel:

$$M_s(F') = \sigma\left(f^{7\times7}\left([z^s_{avg}; z^s_{max}]\right)\right) \tag{5}$$

The final enhanced output is $F'' = M_s\,(F')\odot F'$. This non-linear filtering process effectively suppresses the activation of non-target regions, achieving a robust "Focus" on tiny object features. As explicitly illustrated in the detailed architecture of Fig. 3, the LFE module rigorously follows this sequential processing paradigm: the top branch first extracts global channel interdependencies via the shared MLP to recalibrate features, while the bottom branch subsequently refines local spatial responses using large-kernel pooling and convolution. The synergy of these two branches dynamically accomplishes adaptive filtration of complex environmental noise.

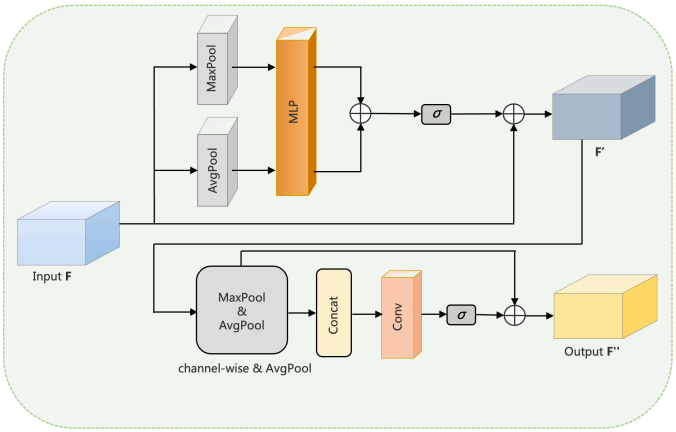

Fig. 3. The detailed structure of the proposed Local Feature Enhancement (LFE) module. It adopts a sequential attention structure, consisting of a Channel Attention Module (top) and a Spatial Attention Module (bottom), to adaptively refine features.

3.3 Content-aware Feature Reassembly (CFR) Module

Standard nearest-neighbor or bilinear interpolations compute solely based on relative spatial coordinates, leading to blurred geometric boundaries for tiny objects. The CFR module achieves lossless reconstruction of fine-grained semantics by generating local reassembly kernels dependent on image content.

Let the input feature map be $X \in \mathbb{R}^{C\times H\times W}$, and the target upsampling ratio be s. Initially, to mitigate computational complexity, we compress the channel dimension to C_m utilizing pointwise convolution, yielding $X_c \in Conv_{1\times1}(X)$. A content encoder ψ_{enc} then projects X_c into the raw reassembly kernel tensor $W' \in \mathbb{R}^{(s^2\cdot k_{up}^2)\times H\times W}$, where k_{up} is the reassembly receptive field.

Following spatial reshaping (Pixel Shuffle), for any target position $l' = (x', y')$ in the upsampled feature map $Y \in \mathbb{R}^{C\times sH\times sW}$, we perform local normalization on its corresponding reassembly kernel via a Softmax operation along the channel dimension to ensure the weights sum to 1:

$$W_{x',y',k} = \frac{\exp(W'_{x',y',k})}{\sum_{j=1}^{k_{up}^2} \exp(W'_{x',y',j})}, \quad \forall k \in \{1, \ldots, k_{up}^2\} \tag{6}$$

During the feature reassembly phase, we map the target position (x', y') back to the center coordinates $(x, y) = (\lfloor x'/s \rfloor, \lfloor y'/s \rfloor)$ in the source feature map. Let the reassembly radius be $r = \lfloor k_{up}/2 \rfloor$, the pixel value of the reconstructed feature Υ at (x', y') is strictly defined by an explicit 2D weighted summation:

$$\Upsilon(x', y') = \sum_{u=-r}^{r} \sum_{v=-r}^{r} W_{x',y',(u,v)} \cdot X(x+u, y+v) \tag{7}$$

where $W_{x',y',(u,v)}$ is the dynamically adaptive weight at position (x', y') with respect to the neighborhood offset (u, v). This paradigm fundamentally eradicates the semantic misalignment issues intrinsic to standard upsampling. As visually depicted in the micro-architecture of Fig. 4, this module operates through two synergistic branches: the top kernel prediction branch encodes local semantic content to output spatially normalized reassembly weights, while the bottom feature reassembly branch directly utilizes these dynamic weights to perform precise spatial aggregation on the low-resolution input, ultimately yielding the high-fidelity upsampled feature.

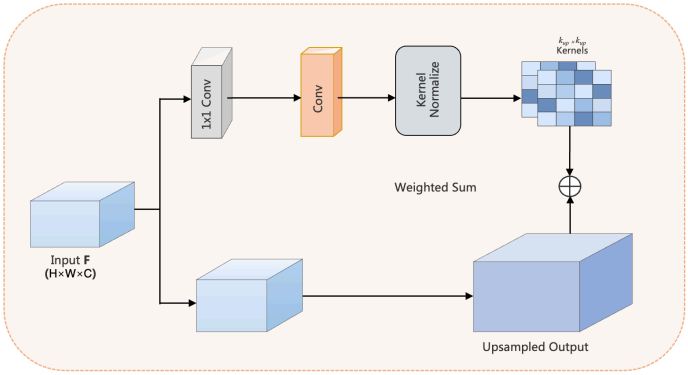

Fig. 4. The detailed structure of the proposed Content-aware Feature Reassembly (CFR) module. It comprises a kernel prediction branch (top) and a feature reassembly branch (bottom) to achieve high-quality content-aware upsampling.

4 Experiment

In this section, we conduct comprehensive empirical evaluations on the authoritative VisDrone2019 benchmark to validate the efficacy of the proposed FR-YOLO framework. We first detail the dataset characteristics, evaluation metrics, and implementation specifics in Section 4.1. Subsequently, we present a quantitative comparison against state-of-the-art (SOTA) lightweight methods in Section 4.2. Finally, we dissect the independent contributions and robustness of the proposed modules through rigorous ablation studies (Section 4.3) and qualitative visualization (Section 4.4).

4.1 Dataset, Evaluation Metrics, and Implementation Details

(1)Dataset and Evaluation Metrics

We evaluate our method on the VisDrone2019 benchmark, universally acknowledged as one of the most challenging large-scale datasets for UAV-based object detection. It comprises 10,209 high-resolution images captured by drones across diverse altitudes, viewing angles, and weather conditions, annotated with ten categories of common ground targets (e.g., pedestrians, vehicles, and cyclists). Specifically, the fundamental challenges of VisDrone2019 stem from the extreme density of tiny objects (often smaller than 32×32 pixels), severe mutual occlusions, and drastic illumination variations. To intuitively illustrate the detection difficulties inherent in these extreme scenarios, along with the significant advantages of our method over the baseline, a representative qualitative visual comparison is provided in Fig. 5 (with detailed analysis deferred to Section 4.4).

For evaluation metrics, we strictly adhere to the standard MS COCO evaluation protocol. We report the Average Precision (AP, averaged over $IoU \in [0.50:0:95]$), $AP_{50}(AP\ at\ IoU = 0.50)$, and $Precision$ as the primary accuracy indicators. Additionally, the number of parameters (Params) and Floating Point Operations (FLOPs) are adopted to holistically assess model computational complexity and deployment friendliness on edge devices.

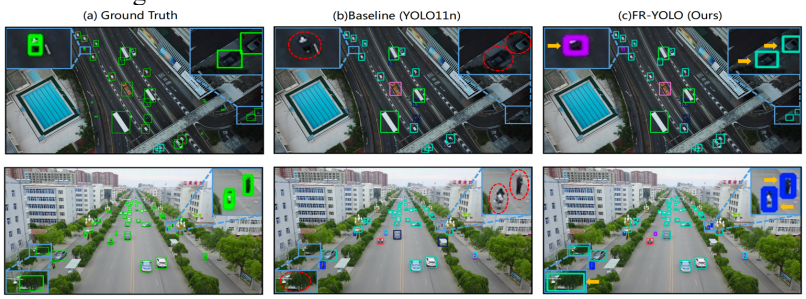

Fig. 5. Visual comparison on VisDrone2019. (a) Ground truth. (b) Baseline misses tiny, occluded objects (red circles) due to noise and misalignment. (c) FR-YOLO accurately localizes them (yellow arrows).

(2)Implementation Details

All experiments were implemented using the PyTorch deep learning framework on a high-performance workstation equipped with a single NVIDIA A100 GPU (40GB VRAM). During preprocessing, input images were uniformly resized to 640×640 pixels. We employed the Stochastic Gradient Descent (SGD) optimizer to train all models from scratch for 300 epochs. To prevent the distortion of tiny objects in the late training stage, Mosaic data augmentation was disabled during the final 10 epochs. The batch size was set to 32 to ensure the stability of the batch normalization statistics. The initial learning rate was set to 0.01, with a momentum of 0.937 and a weight decay of 0.0005. To guarantee absolute fairness in comparison, all evaluated baselines (e.g., YOLOv8n, YOLO11n) were retrained under identical hardware environments and hyperparameter constraints.

4.2 Comparison with State-of-the-Art Methods

Table 1 presents a comprehensive benchmark of FR-YOLO against representative efficient detectors (YOLOv5n/v8n/v10n/11n). To ensure a fair evaluation geared toward edge-device deployment, we strictly focus on the ultra-lightweight domain (Params ≤3.5M), intentionally excluding parameter-heavy models.

The empirical results demonstrate that FR-YOLO achieves an optimal trade-off between detection accuracy and computational efficiency. Specifically, FR-YOLO attains 36.4% AP_{50} and 21.3% AP, representing substantial absolute improvements of 2.5% and 1.5% over the baseline YOLO11n, respectively. Even when compared to the highly robust YOLOv8n, it maintains a 0.8% lead in AP_{50}. Regarding model efficiency, our method exhibits superlative lightweight characteristics: it requires the lowest computational cost (6.4G FLOPs) among all evaluated methods while maintaining a highly compact footprint of 2.53M parameters. Notably, compared to the widely adopted YOLOv8n, FR-YOLO achieves a 15.9% reduction in parameters and a 21.9% decrease in FLOPs while comprehensively dominating in precision. This compellingly validates the architecture's effectiveness in mitigating feature degradation and its immense practical value for deployment on power-constrained edge platforms.

Table 1. Performance comparison on the VisDrone2019 dataset.

Method	Higher is better			Lower is better	
	AP_{50}	AP	Precision	Params	FLOPs
YOLOv5n	33.7	19.5	42.8	**2.51M**	7.2G
YOLOv8n	35.6	20.6	47.5	3.01M	8.2G
YOLOv10n	34.1	19.7	44.4	2.71M	8.4G
YOLO11n	33.9	19.8	44.8	2.59M	6.5G
FR-YOLO(Ours)	**36.4**	**21.3**	**48.3**	2.53M	**6.4G**

4.3 Ablation Studies

To quantitatively decouple and validate the independent contributions of the Local Feature Enhancement (LFE) and Content-aware Feature Reassembly (CFR) modules, we conducted a rigorous ablation study on the YOLO11n baseline. The detailed results are presented in Table 2.

The quantitative analysis reveals a conclusion of high academic value: neither module can achieve positive performance gains in isolation. Applying the LFE module alone results in a marginal performance drop (to 19.4% AP), which mathematically corroborates that isolated high-frequency spatial filtering inevitably attenuates the local contextual information required for localization while suppressing background noise. Similarly, the standalone application of the CFR module yields negligible gains (AP stagnates at 19.8%), as dynamic reassembly kernels tend to fit and reconstruct noise features rather than target features when the input is severely contaminated by environmental interference.

The pivotal performance leap occurs exclusively when both modules are jointly integrated, propelling the overall performance to 21.3% AP and 36.4% AP_{50}, alongside a marked surge in Precision to 48.3%. This phenomenon strongly corroborates an inseparable "Synergistic Coupling" effect: the LFE module acts as a pre-filter providing

high-purity, noise-free salient features, establishing a high-quality data foundation for the subsequent CFR module to correctly predict dynamic reassembly kernels. This validates that the structural integrity of the "Focus-and-Reconstruct" mechanism is indispensable for achieving superior detection.

Table 2. Ablation study of the proposed modules.

Method	LFE	CFR	AP_{50}	AP	Precision
YOLO11n (Baseline)			33.9	19.8	44.8
YOLO11n + LFE	√		33.6	19.4	42.4
YOLO11n + CFR		√	34.2	19.8	45.8
FR-YOLO (Ours)	√	√	**36.4**	**21.3**	**48.3**

4.4 Qualitative Analysis

To intuitively substantiate the spatial robustness of FR-YOLO, Fig. 5 (presented earlier) demonstrates the qualitative visual comparisons against the YOLO11n baseline on the VisDrone2019 test set.

In the highly challenging illumination and shadow scenarios (row 1), the baseline suffers severe feature degradation due to its inability to resist the high-frequency interference from tree shadows and road reflections, completely failing to detect pedestrians and non-motorized vehicles concealed beneath the shadows. In stark contrast, FR-YOLO precisely penetrates the interference to localize these elusive targets. This is directly attributable to the LFE module, which effectively filters non-target region responses and amplifies the saliency of tiny targets via its spatial attention mechanism.

Furthermore, in highly crowded and severely occluded top-down scenes (row 2), the baseline exhibits obvious localization drift and boundary adhesion (within red dashed circles)—a typical semantic spatial misalignment issue induced by bilinear or nearest-neighbor upsampling. Driven by the CFR module, FR-YOLO successfully rectifies the spatial misalignment in the feature dimension and losslessly recovers the fine-grained geometric boundaries of the targets. Consequently, our model flawlessly disentangles the boundaries of overlapping targets in crowded environments (as marked by yellow arrows). These compelling visual proofs flawlessly cross-validate the quantitative enhancements in Table 1, rigorously confirming FR-YOLO's superior capability in navigating complex geometric deformations and feature degradation inherent in UAV imagery.

5 Conclusion

This paper addressed the challenge of detecting tiny objects in UAV imagery under strict computational constraints. By proposing FR-YOLO, we established a novel "Focus-and-Reconstruct" paradigm that effectively bridges the gap between detection accuracy and inference efficiency. Our method demonstrates that the synergistic integration of spatial attention (LFE) and content-aware reassembly (CFR) is a robust solution for mitigating feature degradation in deep networks. Results on VisDrone2019 confirm FR-YOLO outperforms SOTA lightweights (e.g., YOLOv8n) with 22% lower FLOPs.

While specialized models exist, ours offers a practical base-line for industrial applications. These findings validate FR-YOLO as a compelling solution for real-time aerial surveillance on resource-limited edge platforms. Moving forward, our research will extend to hardware-aware optimization techniques, such as network pruning and quantization, to further maximize deployment efficiency in real-world scenarios.

References

1. Telikani, A., Sarkar, A., Du, B., Shen, J.: Machine Learning for UAV-Aided ITS: A Review With Comparative Study. IEEE Trans. Intell. Transport. Syst. 25(11), 15388–15406 (2024)
2. Alam, M.M., Moh, S.: Joint Optimization of Trajectory Control, Task Offloading, and Resource Allocation in Air-Ground Integrated Networks. IEEE Internet Things J. 11(13), 24273–24288 (2024)
3. Yuan, Z., et al.: Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism. Remote Sensing 16(22), 4265 (2024)
4. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object Detection in 20 Years: A Survey. Proc. IEEE 111(3), 257–276 (2023)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. IEEE, Las Vegas, NV, USA (2016)
7. eight Drone Detector for Resource-Constrained Cameras. IEEE Internet Things J. 11(6), 11046–11057 (2024)
8. Xiong, X., et al.: Adaptive Feature Fusion and Improved Attention Mechanism-Based Small Object Detection for UAV Target Tracking. IEEE Internet Things J. 11(12), 21239–21249 (2024)
9. Bakirci, M.: Performance evaluation of low-power and lightweight object detectors for real-time monitoring in resource-constrained drone systems. Engineering Applications of Artificial Intelligence 159, 111775 (2025)
10. Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., Liang, R.: AFPN: Asymptotic Feature Pyramid Network for Object Detection. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2184–2189. IEEE, Honolulu, Oahu, HI, USA (2023)
11. Terven, J., Córdova-Esparza, D.-M., Romero-González, J.-A.: A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. MAKE 5(4), 1680–1716 (2023)
12. Wu, X., Sahoo, D., Hoi, S.C.H.: Recent advances in deep learning for object detection. Neurocomputing 396, 39–64 (2020)
13. Du, Z., Liang, Y.: Object Detection of Remote Sensing Image Based on Multi-Scale Feature Fusion and Attention Mechanism. IEEE Access 12, 8619–8632 (2024)
14. Li, Y., et al.: Lightweight Object Detection Networks for UAV Aerial Images Based on YOLO. Chinese J. Elect. 33(4), 997–1009 (2024)
15. Cao, Q., Chen, Y., Ma, C., Yang, X.: Few-Shot Rotation-Invariant Aerial Image Semantic Segmentation. IEEE Trans. Geosci. Remote Sensing 62, 1–13 (2024)

16. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, LNCS, vol. 11211, pp. 3–19. Springer International Publishing, Cham (2018)
17. Jia, N., Zhen, M., Lv, D., Liu, X., Wei, Z.: LEFE-Net: Lightweight Edge Feature Enhancement for Salient Object Detection in UAV Real-Time Systems. IEEE Internet Things J., 1–1 (2025)
18. Xie, X., You, Z.-H., Chen, S.-B., Huang, L.-L., Tang, J., Luo, B.: Feature Enhancement and Alignment for Oriented Object Detection. IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing 17, 778–787 (2024)
19. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: CARAFE: Content-Aware ReAssembly of FEatures. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3007–3016. IEEE, Seoul, Korea (South) (2019)
20. Lu, Y., Sun, M.: SSE-YOLO: Efficient UAV Target Detection With Less Parameters and High Accuracy. Computer Science and Mathematics (2024)
21. Wang, C.-Y., Liao, H.-Y.M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., Yeh, I.-H.: CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1571–1580. IEEE, Seattle, WA, USA (2020)