SF-YOLO: Scale-Aware and Frequency-Enhanced YOLO for Real-Time Aerial Object Detection

1st Xiaoan Bao
Zhejiang Sci-Tech University
Hangzhou, China
baoxiaoan@zstu.edu.cn

2nd Cheng Xu
Zhejiang Sci-Tech University
Hangzhou, China
958996432@qq.com

3rd Haidi Xu
Zhejiang Sci-Tech University
Hangzhou, China
x1078303266@gmail.com

4th Na Zhang
Zhejiang Sci-Tech University
Hangzhou, China
103185968@qq.com

5th Biao Wu
Zhejiang Sci-Tech University
Hangzhou, China
biaowuzg@zstu.edu.cn

6th Qingqi Zhang*
Hangzhou Institute of Medicine
Chinese Academy of Sciences
Hangzhou, China
zhangtsingtsi@gmail.com

***Abstract*—The rapid development of Unmanned Aerial Vehicle (UAV) technology has established object detection in aerial imagery as a core supporting technology in fields such as urban surveillance and disaster relief. However, constrained by extreme scale variations, the irreversible loss of tiny object features, and the stringent latency constraints of onboard edge devices, real-time object detection in this domain still faces severe challenges. Although FBRT-YOLO has achieved excellent inference speeds through architectural pruning, its aggressive downsampling strategy trades efficiency at the cost of tiny object perception, leading to severe missed detections. To address this challenge, we propose a Scale-aware and Frequency-enhanced detector, SF-YOLO, aiming to break through the bottleneck of tiny object localization precision while maintaining real-time processing capabilities. First, we introduce a Selective Kernel Downsampling (SKD) module, which effectively circumvents feature aliasing and information loss caused by fixed convolutions by incorporating a dynamic receptive field mechanism in the shallow feature extraction stage. Second, addressing the resolution bottleneck, we design a lightweight High-Frequency Perception (HFPlite) module that utilizes high-pass filtering priors and a dual-path attention mechanism to efficiently recover fine-grained target textures while suppressing background noise. Finally, we construct a Weighted Inner-Wasserstein IoU (WIW-IoU) loss function, which fundamentally resolves the gradient vanishing problem caused by the sensitivity of tiny objects to positional deviations by synergizing distribution metrics with geometric constraints. Experimental results demonstrate that, compared with the baseline model FBRT-YOLO-S, SF-YOLO-S significantly improves mAP_{50} by 3.4% and 3.8% on the VisDrone2018 and AI-TOD datasets respectively, with a 6.9% reduction in parameters, successfully achieving the optimal balance between detection performance and inference speed.**

***Index Terms*—Object Detection, Aerial Imagery, Convolutional Neural Networks, Tiny Object Detection, Real-time Processing.**

I. Introduction

With the proliferation of Unmanned Aerial Vehicle (UAV) technology, object detection in aerial imagery has demonstrated immense application value in fields such as urban surveillance, disaster relief, and geographic mapping. However, compared to natural scenes (e.g., COCO), aerial scenarios face more severe challenges. First, extreme scale variations, where fluctuations in UAV flight altitude lead to significant differences in target sizes. Second, the dense distribution of tiny objects (often smaller than 16×16 pixels), whose geometric structures and textural information are highly susceptible to loss during the layer-wise downsampling of deep Convolutional Neural Networks (CNNs) [1], [2]. Furthermore, the limited computational power of onboard devices requires that algorithms satisfy strict real-time constraints while maintaining high precision.

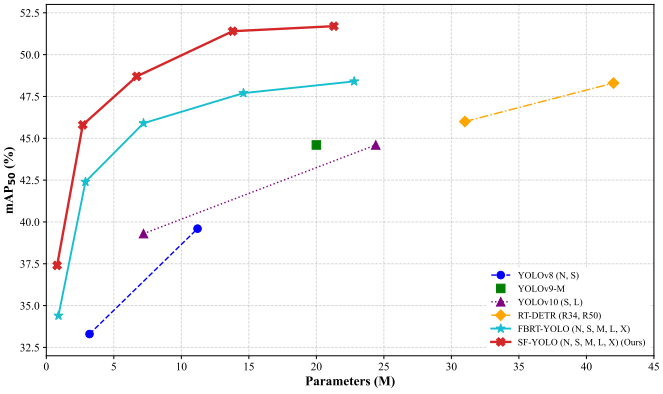

Fig. 1. Comparison of model parameters and mAP_{50}

To address these difficulties, general-purpose real-time detectors, such as the YOLO series [3], SSD [4], and lightweight architectures like MobileNet [5], have been widely improved to adapt to aerial scenarios. Recently, FBRT-YOLO [6] achieved an excellent balance between precision and speed by introducing Feature Complementary Mapping (FCM) and Multi-Kernel Perception (MKP) modules to construct an efficient lightweight architecture. However, in its pursuit of extreme inference speed, FBRT-YOLO aggressive de-redundancy strategy reveals two key flaws when processing tiny objects.

First is "late perception." Although FBRT-YOLO uses MKP

*Corresponding author: Qingqi Zhang (zhangtsingtsi@gmail.com).

to enhance multi-scale expression, this module is only deployed at the end of the backbone network. Feature loss of tiny objects usually occurs during the early downsampling stages of the network. Relying on fixed standard convolution kernels for early downsampling causes the structural information of tiny objects to be aliased or eroded before it can be transmitted to the deep MKP, resulting in irreversible feature damage.

Second is the "resolution bottleneck." To reduce computational load, FBRT-YOLO removes the high-resolution P3 detection head. Although this strategy improves inference speed, it also blocks the path for the network to obtain shallow high-frequency details, severely limiting the recall rate of tiny objects.

In response to these limitations, we propose a Scale-aware and Frequency-enhanced detector, SF-YOLO, which retains the real-time advantages of FBRT-YOLO. Specific improvements include: First, addressing "late perception," we introduce the Selective Kernel Downsampling (SKD) module at the front end of the backbone to replace standard downsampling. SKD utilizes multi-branch depthwise convolutions and a parameter-free attention mechanism to dynamically adjust the receptive field at the early stage of feature extraction, thereby effectively preserving tiny object features and adapting to extreme scale variations. Second, to break the "resolution bottleneck," we reconstruct the high-resolution P3 branch and embed a lightweight High-Frequency Perception (HFPlite) module. This module leverages high-pass filtering priors to recover the edge textures of tiny objects and suppress shallow background noise. Finally, addressing the extreme sensitivity of tiny objects to positional deviations, we design the Weighted Inner-Wasserstein (WIW-IoU) loss. This function creatively integrates the distribution metric advantages of Normalized Wasserstein Distance (NWD) [19] with the geometric convergence advantages of Inner-IoU, achieving dual improvements in robustness and convergence speed.

Extensive experiments on the VisDrone2018 [7] and AI-TOD [8] benchmarks corroborate the efficacy of our approach. As depicted in Fig. 1, which illustrates a comparative analysis of prevailing models on the VisDrone2018 validation set, SF-YOLO significantly boosts tiny object detection performance in aerial imagery while adhering to a lightweight design. The primary contributions are summarized as follows:

- We propose a Selective Kernel Downsampling (SKD) module that adaptively adjusts the receptive field and aggregation weights based on input features. This module reduces the loss of tiny object features caused by rigid downsampling.
- We reconstruct the high-resolution P3 detection branch and design a lightweight High-Frequency Perception (HFPlite) module. This module effectively suppresses background noise and recovers fine-grained geometric details of tiny objects, addressing the resolution bottleneck problem.
- We design a novel WIW-IoU loss function that synergizes the distribution-based NWD and the geometry-based Inner-IoU. This hybrid metric resolves the gradient vanishing problem triggered by the extreme sensitivity of tiny objects to positional deviations and accelerates training convergence.

II. Related Work

A. Real-time Object Detection in Aerial Imagery

For the past decade, single-stage detectors, epitomized by the YOLO series from Redmon et al. [3], have commanded the domain of general-purpose real-time detection owing to their superior inference efficiency. This evolution has been further accelerated by recent state-of-the-art models such as YOLOv8 [9], YOLOv9 [10], and YOLOv10 [11], as well as the transformer-based RT-DETR [13], which continue to set new benchmarks for the speed-accuracy trade-off. To align these generic architectures with the intricacies of aerial imagery, extensive adaptations have been proposed. A case in point is TPH-YOLOv5 by Zhu et al. [14], which substantially augments global context capture in dense scenarios by integrating Transformer prediction heads. More recently, the stringent computational constraints of edge devices have propelled lightweight architectures to the forefront of research. Exemplifying this trajectory, FBRT-YOLO, proposed by Xiao et al. [6], achieves rapid inference velocities by pruning redundant detection heads. Nevertheless, such aggressive downsampling, while prioritizing efficiency, precipitates irreversible feature degradation for tiny objects during shallow feature extraction, thereby imposing a rigid ceiling on detection accuracy.

B. Multi-Scale Perception Strategies

Since the inception of FPN by Lin et al. [15], feature pyramids have established themselves as the prevailing paradigm for tackling multi-scale issues. To circumvent the receptive field constraints inherent in fixed convolution kernels, SKNet by Li et al. [16] and the MKP unit by Xiao et al. incorporated dynamic mechanisms that leverage attention to adaptively modulate kernel sizes. Nevertheless, contemporary dynamic perception modules are predominantly stationed in deep network layers to aggregate high-level semantics. Conversely, shallow downsampling continues to rely principally on standard fixed-stride convolutions. This structural bias causes the geometric integrity of tiny objects to suffer aliasing or erosion prior to reaching deeper layers, ensuring that the challenge of early feature corrosion remains effectively unmitigated.

C. Optimization for Bounding Box Regression

Nascent endeavors were primarily dedicated to refining the fidelity of geometric metrics. By integrating center-point distance and angular constraints, CIoU proposed by Zheng et al. [17] and SIoU by Gevorgyan [18] substantially bolstered regression convergence. Yet, these IoU-based indices manifest an acute hypersensitivity to the positional deviations of tiny objects. To remedy this pathology, Wang et al. introduced NWD [19], which innovatively models objects as Gaussian distributions, effectively circumventing the gradient vanishing problem in non-overlapping regions. However, notwithstanding the robustness conferred by NWD, its deficit in rigorous

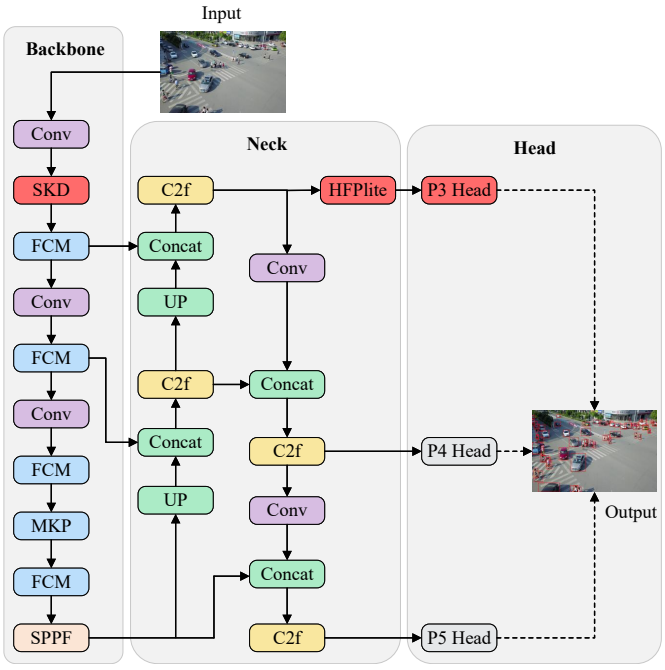

Fig. 2. Model structure diagram of SF-YOLO

geometric constraints often precipitates convergence bottlenecks, particularly when juxtaposed with emerging acceleration methodologies such as Inner-IoU by Zhang et al. [20].

III. Methods

In this section, we first present an overview of the proposed SF-YOLO architecture. Subsequently, we delineate three pivotal innovative components: the Selective Kernel Downsampling (SKD) module, designed to mitigate early perception latency; the lightweight High-Frequency Perception (HFPlite) module, tailored to surmount the resolution bottleneck; and the Weighted Inner-Wasserstein (WIW-IoU) loss, formulated to optimize tiny object regression.

A. Overall Architecture of SF-YOLO

SF-YOLO is constructed upon the lightweight foundation of FBRT-YOLO, designed to bridge the performance gap in tiny object detection while inheriting the baseline's exceptional inference speed. Although FBRT-YOLO achieves high computational efficiency through feature complementarity and multi-kernel perception, its aggressive downsampling strategy and the removal of the P3 detection head cause tiny object features to be lost during early encoding. As illustrated in Fig. 2, we propose a series of strategic modifications to resolve this conflict while preserving the efficient decoupled head and cross-layer connectivity.

To address early perception latency in the backbone, we specifically replace the standard convolution in the initial downsampling stage with the SKD module. This integration aims to introduce a dynamic receptive field mechanism, enabling the network to adapt to the drastic scale variations inherent in aerial imagery at the very beginning of feature extraction, thereby preventing the premature loss of tiny features.

Furthermore, to overcome the resolution bottleneck, we reconstruct the neck architecture by reinstating the high-resolution P3 branch, previously omitted in the original version and embedding the lightweight HFPlite module within the feature fusion path. This design leverages high-pass filtering priors to effectively recover fine-grained texture details while actively suppressing background noise interference common in shallow feature maps.

Finally, addressing the instability of regression, we optimize the training phase by adopting the WIW-IoU loss function to replace the conventional bounding box loss. This substitution provides a smooth optimization manifold for tiny objects that are hypersensitive to positional deviations, effectively resolving the gradient vanishing problem associated with traditional IoU metrics.

B. SKD Module

Standard CNN backbones predominantly rely on fixed-stride convolutions (e.g., 3×3 with stride 2) for initial spatial downsampling. While effective for uniform objects, this static receptive field is ill-equipped to handle the significant scale variations inherent in aerial imagery. To address this limitation, we introduce the SKD module, as shown in Fig. 3.

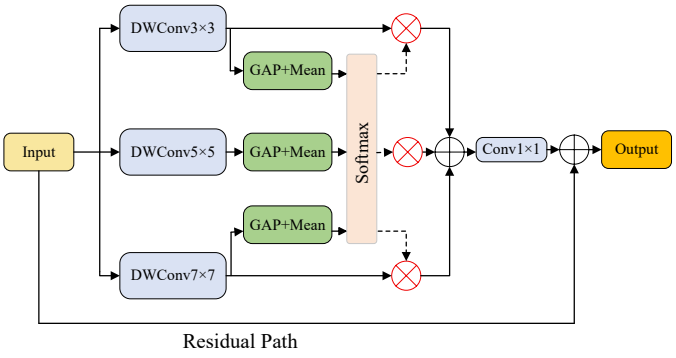

Fig. 3. The structure of the SKD module, the solid lines denote the data flow of feature maps, while the dashed lines represent the control flow of attention weights.

The SKD module replaces the standard downsampling layer in the first stage of the backbone. Distinct from the original SKNet which utilizes channel-wise attention, our design focuses on lightweight, branch-level adaptability. The module comprises three integrated components:

First, let $X \in \mathbb{R}^{C \times H \times W}$ denote the input feature map. We apply a projection layer F_{proj} to adjust spatial resolution and channel dimensions. The projected features are subsequently processed by three parallel Depth-wise Convolution (DWConv) branches with distinct kernel sizes $k \in \{3, 5, 7\}$:

$$U_k = \text{DWConv}_k(F_{proj}(X)), \quad k \in \{3, 5, 7\} \tag{1}$$

where U_k denotes the feature response generated by the branch with kernel size k, and F_{proj} represents a convolution with stride s (typically $s = 2$ for downsampling). This design allows the network to capture diverse spatial contexts with minimal parameter overhead.

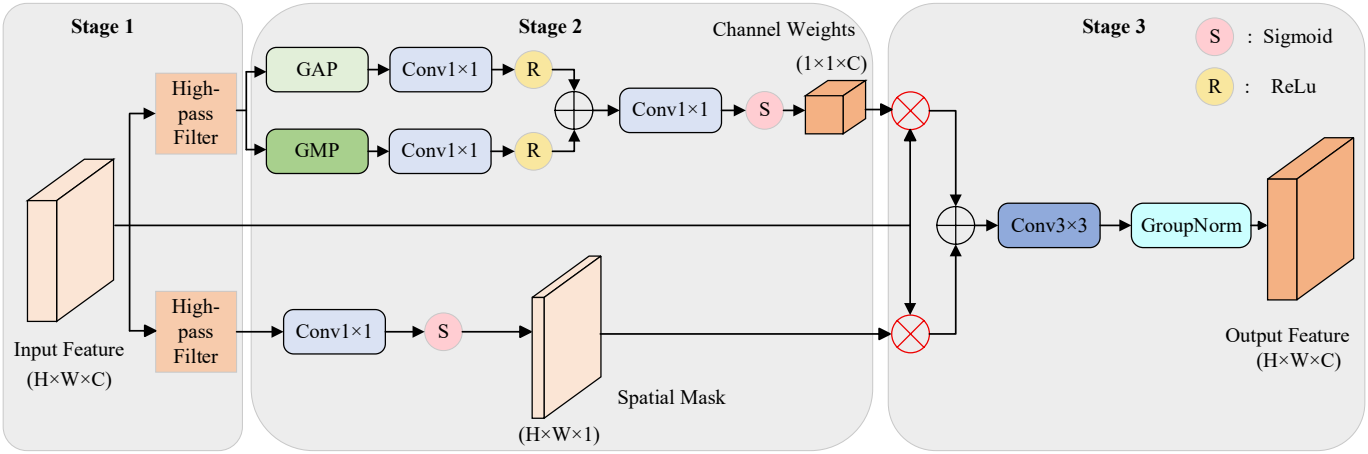

Fig. 4. The structure of the HFPlite module

Second, to ensure computational efficiency, we propose a scalar attention mechanism as a substitute for computationally intensive channel-wise operations. For each branch k, we compress the feature map U_k into a single scalar descriptor v_k via Global Average Pooling (GAP) followed by a channel-averaging operation:

$$v_k = \frac{1}{C}\sum_{c=1}^{C} \text{GAP}(U_k^c) \tag{2}$$

where U_k^c represents the channel with index c of feature map U_k, and C is the number of channels. These descriptors are subsequently stacked and normalized via a Softmax function to generate adaptive weights $w \in \mathbb{R}^3$, satisfying $\sum w_k = 1$. This mechanism prioritizes the most appropriate kernel size based on global image context.

Finally, the calibrated multi-scale features are aggregated via weighted summation. A fusion layer integrates cross-channel information, and a residual connection is introduced to maintain gradient flow:

$$Y = F_{fuse}\left(\sum_{k\in\{3,5,7\}} w_k \cdot U_k\right) + F_{short}(X) \tag{3}$$

Here, Y denotes the final output tensor. F_{fuse} is a pointwise convolution for channel fusion, and F_{short} represents the projection layer in the residual path to ensure dimensional alignment. By integrating the SKD module, the backbone achieves dynamic scale awareness at the earliest stage of computation.

C. HFPlite Module and P3 Branch

To surmount the "resolution bottleneck" inherent in tiny object detection, we reinstate the high-resolution P3 detection branch. However, while shallow features preserve spatial details, they are often plagued by significant background noise (e.g., complex environmental textures), where direct fusion can easily lead to erroneous detections. To resolve this conflict, we embed the lightweight HFPlite module within the P3 branch, utilizing a high-pass filtering prior to guide feature enhancement. As illustrated in Fig. 4, the processing flow of this module is formalized into three stages:

First, unlike approaches that rely solely on learnable convolutions, we introduce a fixed Laplacian operator to preprocess the input feature $X \in \mathbb{R}^{C\times H\times W}$. This step utilizes a second-order differential operator to explicitly separate the high-frequency components of the image, generating the high-frequency response map F_h:

$$F_h = X * K_{Lap} \tag{4}$$

where $*$ denotes the convolution operation, and K_{Lap} is the predefined Laplacian kernel. This fixed prior effectively suppresses low-frequency background redundancy, providing explicit "edge guidance" for the subsequent attention mechanism.

Second, to further discriminate valid tiny object textures from redundant background within F_h, we construct two parallel attention paths to generate channel weights W_{ch} and spatial masks W_{sp}. In the channel dimension, aiming to adaptively select feature responses rich in high-frequency details, we first perform GAP and GMP on F_h in parallel to aggregate global statistics. These pooled features are individually transformed by a grouped convolution $\mathcal{F}_{group}$ with ReLU activation δ, and then summed to form the aggregated feature F_{agg}. Finally, F_{agg} is mapped to the channel weights W_{ch} via a second grouped convolution followed by a Sigmoid function σ. Simultaneously, in the spatial dimension, to precisely localize regions with dense high-frequency textures (i.e., potential tiny objects), we project F_h directly through a pointwise convolution F_{point} to generate the spatial mask W_{sp}. Specifically, the calculations for W_{ch} and W_{sp} are defined as follows:

$$F_{agg} = \delta\left(\mathcal{F}_{group}(\text{GAP}(F_h))\right) + \delta\left(\mathcal{F}_{group}(\text{GMP}(F_h))\right) \tag{5}$$

$$W_{ch} = \sigma(F_{group}(F_{agg})) \tag{6}$$

$$W_{sp} = \sigma(F_{point}(F_h)) \tag{7}$$

Finally, the original input feature X is separately modulated by the attention weights from these two paths, and then aligned and fused via a final convolution layer. The final output feature Y is calculated as:

$$Y = F_{out}(X \odot W_{ch} + X \odot W_{sp}) \tag{8}$$

where $\odot$ denotes element-wise multiplication, and F_{out} represents the post-processing block consisting of a 3×3 convolution and Group Normalization to fuse the modulated features. Finally, Y serves as the enhanced feature output to the detection head. Through this design, HFPlite significantly improves the model's perception of tiny textured objects in complex backgrounds.

D. Weighted Inner-Wasserstein Loss

Bounding box regression in tiny object detection faces a critical trade-off between convergence speed and robustness. Traditional IoU-based metrics (e.g., CIoU [17], SIoU [18]) exhibit hypersensitivity to positional deviations, leading to gradient vanishing in non-overlapping cases. While NWD [19] mitigates this by modeling bounding boxes as Gaussian distributions, it suffers from slow convergence. Conversely, Inner-IoU [20] accelerates regression via auxiliary box scaling but lacks distributional robustness for tiny targets.

To synthesize the strengths of these approaches, we propose WIW-IoU loss function. First, following the NWD paradigm, we model the predicted box B_p and ground truth box B_{gt} as 2D Gaussian distributions $\mathcal{N}_p$ and $\mathcal{N}_{\text{gt}}$, respectively, and quantify their geometric affinity using the normalized Wasserstein distance:

$$\text{NWD}(B_p, B_{\text{gt}}) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_p, \mathcal{N}_{\text{gt}})}}{C}\right) \tag{9}$$

where C is a constant. This probabilistic formulation ensures a continuous gradient flow even in the absence of spatial overlap, effectively resolving the gradient vanishing problem of traditional IoU in non-overlapping scenarios.

Building on this, WIW-IoU establishes a dynamic optimization trajectory through a weighted formulation:

$$L_{WIW-IoU} = L_{Inner-IoU} + \lambda(1 - NWD) \tag{10}$$

This mechanism leverages complementary gradient advantages: during the early training stage (non-overlapping), the NWD term dominates the gradient flow to guide the predicted box rapidly towards the target region; once overlap occurs, the Inner-IoU term takes precedence, providing high-sensitivity geometric gradients for precise boundary alignment. This strategy effectively circumvents the risk of gradient vanishing associated with single geometric metrics, achieving a dual improvement in convergence speed and localization accuracy.

IV. Experiment and Result Analysis

A. Datasets and Implementation Details

We conduct extensive experiments on two aerial object detection benchmarks: VisDrone2018 [7] and AI-TOD [8]. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. Our network is trained for 300 epochs using the stochastic gradient descent (SGD) optimizer with a momentum of 0.937, a weight decay of 0.0005, a batch size of 4, and an initial learning rate of 0.01.

B. Evaluation Metrics

To comprehensively evaluate the detection performance and computational efficiency of the SF-YOLO model, we employ Precision, Recall, Mean Average Precision (mAP), Parameters (Params), and Floating Point Operations (FLOPs) as evaluation metrics.

Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive, while Recall measures the proportion of actual positive samples correctly predicted by the model. Their calculations are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

where TP, FP, and FN represent the number of True Positives, False Positives, and False Negatives, respectively.

As the authoritative evaluation standard for object detection, mAP is obtained by calculating the mean of Average Precision (AP) across all classes. For a single class, AP is defined as the integral area under the Precision-Recall (P-R) curve:

$$AP = \int_0^1 P(R)dR \tag{13}$$

The mAP over all classes is calculated as:

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{14}$$

where C is the number of classes. In this experiment, we focus on mAP_{50}, $\text{AP}_{50:95}$ (the mean at IoU thresholds from 0.5 to 0.95), and AP_{75} (precision at IoU threshold 0.75) to evaluate localization performance under different overlap degrees.

Params represents the total number of learnable parameters, reflecting spatial complexity and memory usage. FLOPs denotes the number of floating-point operations for a single inference, directly correlating with the computational load and inference latency of the model.

C. Comparative experiments

As shown in Table I, we comprehensively compare SF-YOLO and its baseline FBRT-YOLO with existing mainstream real-time detectors. The experimental results demonstrate that our method achieves the optimal balance between parameter efficiency and detection accuracy, whether on lightweight small models or high-performance large models. For small models, compared to the baseline FBRT-YOLO-N/S, SF-YOLO-N/S reduces the parameter count by 11.1% and 6.9%, respectively, while significantly improving mAP_{50} by 3% and

TABLE I
COMPARISON OF AP, mAP_{50}, PARAMS AND FPS ON VISDRONE2018 BY USING OUR METHODS WITH DIFFERENT REAL-TIME OBJECT DETECTORS.

Model	AP (%)	mAP_{50} (%)	FLOPs (G)	Params (M)	FPS
YOLOv8-N (2023)	19.6	33.3	8.7	3.2	212
YOLOv8-S (2023)	23.6	39.6	28.6	11.2	165
YOLOv8-M (2023)	27.1	44.4	78.9	25.9	102
YOLOv8-L (2023)	28.4	45.9	165.2	43.7	75
YOLOv8-X (2023)	28.9	46.8	257.8	68.2	58
YOLOv9-M (2024)	27.2	44.6	76.3	20.0	96
YOLOv10-S (2024)	23.8	39.3	21.6	7.2	168
YOLOv10-L (2024)	27.6	44.6	120.3	24.4	80
RT-DETR-R34 (2024)	27.2	46.0	92.0	31.0	108
RT-DETR-R50 (2024)	28.8	48.3	136.0	42.0	82
FBRT-YOLO-N (2025)	20.2	34.4	6.7	0.9	235
FBRT-YOLO-S (2025)	25.9	42.4	22.9	2.9	175
FBRT-YOLO-M (2025)	28.4	45.9	58.7	7.2	115
FBRT-YOLO-L (2025)	29.7	47.7	119.2	14.6	85
FBRT-YOLO-X (2025)	30.1	48.4	185.8	22.8	64
SF-YOLO-N (Ours)	22.3	37.4	10.8	0.8	219
SF-YOLO-S (Ours)	28.1	45.8	30.8	2.7	167
SF-YOLO-M (Ours)	30.2	48.7	67.3	6.7	108
SF-YOLO-L (Ours)	31.8	51.4	140.4	13.8	78
SF-YOLO-X (Ours)	**32.2**	**51.7**	215.3	21.3	60

TABLE II
COMPARISON WITH OTHER STATE-OF-THE-ART DETECTORS ON VISDRONE2018.

Model	AP (%)	mAP_{50} (%)	AP_{75} (%)
DMNet (2020) [22]	29.4	49.3	30.6
QueryDet (2022) [23]	28.3	48.1	28.8
CEASC (2023) [24]	28.7	50.7	28.4
YOLC (2024) [25]	28.9	51.4	28.3
KLDet (2024) [26]	28.5	50.7	28.1
FBRT-YOLO-X (2025) [6]	30.1	48.4	31.7
SF-YOLO-X (Ours)	**32.2**	**51.7**	**33.8**

3.4%, and AP by 2.1% and 2.2%, respectively. For medium models, compared with FBRT-YOLO-M and YOLOv9-M, SF-YOLO-M reduces parameters by 6.9% and 66.5%, respectively, while increasing AP by 1.8% and 3.0%, respectively. For large models, compared to FBRT-YOLO-X, our SF-YOLO-X reduces parameters by 6.6%, improves mAP_{50} by 3.3%, and significantly boosts AP by 2.1%. Furthermore, compared to RT-DETR-R34/R50, SF-YOLO-M/L exhibits fewer parameters, lower GFLOPs, faster detection speeds, and superior detection performance. These experimental results fully prove the superiority of SF-YOLO as a real-time aerial image detector.

As shown in Table II, we present the comparison results of our method with other state-of-the-art methods on VisDrone, indicating that SF-YOLO effectively detects aerial images.

To verify the generalization ability of SF-YOLO on tiny object detection, we conducted supplementary experiments on the AI-TOD dataset. Table III presents the comparison results of SF-YOLO-S with other mainstream algorithms on the AI-TOD dataset, where experimental data indicate that SF-YOLO-S demonstrates significant advantages. Specifically, SF-YOLO-S achieves an AP of 22.4%, representing a 2.2% improvement over the baseline model FBRT-YOLO-S and outperforming high-performance algorithms such as MeNet and KLDet. Regarding parameter count, SF-YOLO-S requires only 2.7 M parameters, ranking as the smallest model among those compared; this constitutes a reduction of approximately 6.9% compared to FBRT-YOLO-S and about 93% compared to Faster-RCNN and MeNet. Furthermore, although the computational cost (30.8 GFLOPs) of SF-YOLO-S is slightly higher than that of the baseline, it achieves a score of 49.6% on the mAP_{50} metric, significantly surpassing the baseline by 3.8%. In summary, SF-YOLO-S effectively enhances the capture capability for tiny objects while maintaining an extremely low parameter count, achieving a favorable balance between model lightweighting and detection accuracy.

TABLE III
COMPARISON ON AI-TOD BY USING OUR METHODS WITH OTHER STATE-OF-THE-ART DETECTORS.

Model	AP (%)	mAP_{50} (%)	FLOPs (G)	Params (M)
Faster-RCNN [21]	11.6	26.9	134.4	41.2
YOLO11-S [12]	20.1	46.0	21.3	9.4
KLDet [26]	19.6	46.4	339.3	32.0
MeNet [27]	20.4	50.0	148.3	42.0
FBRT-YOLO-S [6]	20.2	45.8	22.9	2.9
SF-YOLO-S (Ours)	**22.4**	**49.6**	30.8	**2.7**

D. Ablation Study

To investigate the independent contributions of each core component, we conducted a stepwise ablation study on the

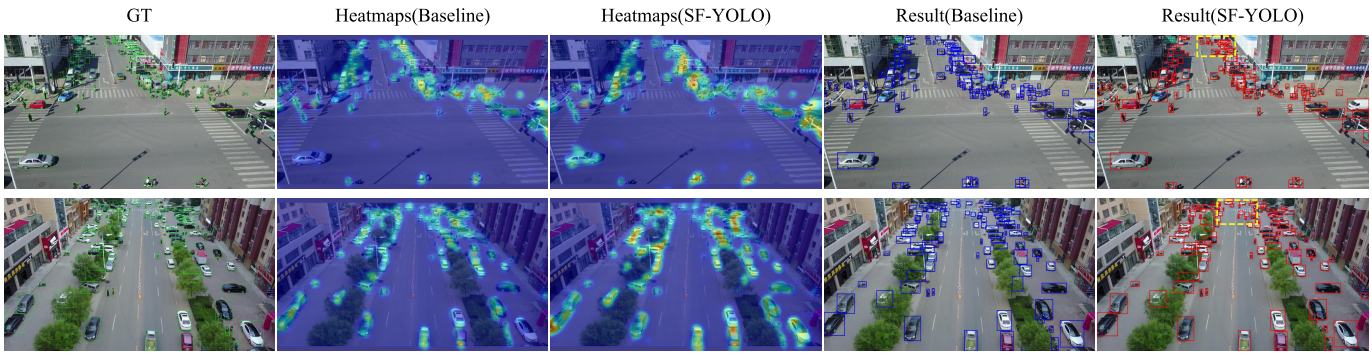

Fig. 5. Visualization comparison of heatmaps and detection results between the baseline and SF-YOLO

VisDrone2018 dataset using SF-YOLO-S as the baseline model. As shown in Table IV, the experimental results present a clear stepwise improvement in performance. First, compared to the original FBRT-YOLO-S baseline, the introduction of the HFPlite module significantly improved AP by 0.9%, strongly proving the necessity of high-pass filtering in recovering texture details blurred by downsampling. Building on this, replacing the standard downsampling layer with the SKD module further increased AP by 0.8%, indicating that the dynamic receptive field mechanism effectively alleviates the feature loss problem caused by extreme scale variations in aerial imagery. Finally, incorporating the WIW-IoU loss function brought the model performance to its peak, increasing AP by 0.5% and mAP_{50} by 0.6%, which directly verifies the superior robustness of WIW-IoU in addressing the gradient vanishing problem in tiny object regression. In summary, the three components constitute a complementary synergistic effect: SKD ensures feature integrity, HFPlite enhances target distinguishability, and WIW-IoU guarantees regression precision. Working together, they achieve a qualitative improvement in tiny object detection performance with a negligible increase in computational burden.

TABLE IV
ABLATION STUDY ON HFPLITE, SKD, AND WIW-IOU MODULE ON VISDRONE2018.

HFPlite	SKD	WIW-IoU	AP (%)	mAP_{50} (%)	Params (M)	FLOPs (G)
			25.9	42.4	2.9	22.9
✓			26.8	44.4	2.6	29.4
✓	✓		27.6	45.2	2.7	30.8
✓	✓	✓	**28.1**	**45.8**	2.7	30.8

E. Parameter Sensitivity Analysis

To explore the optimal hyperparameter setting for the WIW-IoU loss, we conducted a sensitivity analysis on the balancing factor λ in (10). This parameter controls the trade-off between the distribution-based gradient flow (NWD) and the geometry-based boundary regression (Inner-IoU).

We evaluated the performance of SF-YOLO on the VisDrone2018 dataset by varying λ within the range of $[0.5, 2.5]$ with a step size of 0.5. The comparative results in terms of AP, mAP_{50}, and AP_{75} are summarized in Table V.

As illustrated in Table V, the detection performance exhibits an arch-shaped trend, initially increasing and then decreasing as λ rises:

- Under-weighting ($\lambda < 1.0$): With $\lambda = 0.5$, the AP drops to 27.6%. This performance degradation stems from the insufficient NWD weight, which weakens the gradient flow for non-overlapping tiny objects and impedes convergence during early training.
- Over-weighting ($\lambda > 1.5$): Increasing λ to 2.0 or higher negatively impacts performance. An over-emphasized NWD term can suppress the fine-grained geometric constraints offered by Inner-IoU. Since NWD prioritizes distributional overlap over exact boundary alignment, high-precision metrics suffer, as evidenced by AP_{75} dropping from 30.8% to 29.9%.
- Optimal Balance ($\lambda = 1.0$): The model attains its peak performance (AP = 28.1%) at $\lambda = 1.0$. This setting effectively balances the optimization process: NWD guides the early convergence of tiny targets sensitive to positional deviations, enabling Inner-IoU to subsequently refine bounding box precision.

Consequently, we adopt $\lambda = 1.0$ as the default setting for all experiments in this work.

TABLE V
SENSITIVITY ANALYSIS OF THE HYPERPARAMETER λ ON VISDRONE2018.

λ	AP (%)	mAP_{50} (%)	AP_{75} (%)
0.5	27.6	45.1	29.8
1.0	**28.1**	**45.8**	**30.8**
1.5	28.0	45.6	30.5
2.0	27.5	44.9	29.9
2.5	27.1	44.2	29.2

F. Qualitative Analysis and Visualization

To comprehensively evaluate the practical effectiveness of the proposed method, we conducted a visual heatmap analysis to compare the detection capabilities of different models for

tiny objects. Specifically, representative aerial scenes containing dense small targets were selected for visualization on the baseline model and the proposed SF-YOLO, as shown in Fig. 5.

The visualized heatmaps reveal that the baseline is easily distracted by background noise, hindering detection precision. In contrast, SF-YOLO effectively concentrates attention on target regions while suppressing interference, demonstrating the superior feature extraction capability of the SKD module.

In real-world scenarios, such as crowded intersections, the baseline suffers from missed detections. Conversely, SF-YOLO excels in these complex scenes. Specifically, it successfully distinguishes adjacent targets in dense clusters that the baseline merges, validating the efficacy of the HFPlite module in sharpening boundaries. Collectively, these findings substantiate the robustness and high precision of SF-YOLO in complex real-world scenarios.

V. Conclusion

In this work, we presented SF-YOLO, a high-efficiency detector tailored for the challenging domain of aerial imagery. To tackle the inherent issues of extreme scale variation and feature degradation in tiny objects, we reconstructed the detection architecture by integrating three strategic components: the SKD module for dynamic scale adaptation, the HFPlite module for high-frequency feature recovery, and the WIW-IoU loss for robust optimization. These innovations synergistically enhance the network's representation capability while mitigating the gradient vanishing problem. Extensive evaluations on the VisDrone2018 and AI-TOD benchmarks demonstrate that SF-YOLO achieves a superior trade-off between accuracy and latency, outperforming state-of-the-art competitors such as FBRT-YOLO, YOLOv9, and RT-DETR. Our findings suggest that SF-YOLO serves as a robust solution for real-time aerial surveillance. Future work will focus on optimizing the framework for deployment on resource-constrained edge devices.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[3] 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.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[6] Y. Xiao, T. Xu, Y. Xin, and J. Li, "FBRT-YOLO: Faster and better for real-time aerial image detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 8, pp. 8673–8681, 2025.

[7] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and et al., "VisDrone-DET2018: The vision meets drone object detection in image challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 437–468.

[8] 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.

[9] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

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

[11] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and et al., "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[12] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[13] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, and et al., "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16965–16974.

[14] 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.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, and B. Hariharan, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[16] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 510–519.

[17] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12993–13000.

[18] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.

[19] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 294–303, 2021.

[20] H. Zhang, C. Xu, and S. Zhang, "Inner-IoU: More effective intersection over union loss with auxiliary bounding box," *arXiv preprint arXiv:2311.02877*, 2023.

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

[22] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020.

[23] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 13668–13677.

[24] B. Du, Y. Huang, J. Chen, and D. Huang, "Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13435–13444.

[25] C. Liu, G. Gao, Z. Huang, Z. Hu, Q. Liu, and Y. Wang, "YOLC: You only look clusters for tiny object detection in aerial images," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13863–13875, 2024.

[26] Z. Zhou and Y. Zhu, "KLDet: Detecting tiny objects in remote sensing images via Kullback–Leibler divergence," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Art. no. 4703316, 2024.

[27] T. Zhang, X. Zhang, X. Zhu, G. Wang, X. Han, X. Tang, and et al., "Multistage enhancement network for tiny object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.