YOLO-SWIFT: A Small-object Wavelet-Integrated Feature Transform Network for Object Detection in UAV Aerial Imagery

Yanchen Li, Kai Shao
School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
Corresponding author: Kai Shao (e-mail: 18088011330@163.com)

ABSTRACT Accurate detection of small-scale targets in UAV aerial imagery remains challenging due to severe scale variation, background interference, and feature degradation caused by repeated downsampling. To address these issues, this study proposes YOLO-SWIFT, a wavelet-integrated feature transformation network designed for small-object detection. First, a Position-Aware Coordinate Downsampling module is developed to preserve critical spatial information during feature compression through coordinate attention and skip connections. Second, a High-Resolution Feature Aggregation Network is introduced to establish an additional high-resolution detection branch for enhanced cross-scale feature fusion. Third, a Wavelet Bottleneck Enhancement module incorporating multi-level wavelet decomposition and a High-Frequency Retention pathway is designed to improve fine-detail representation while expanding the effective receptive field. Finally, an Adaptive Scale-aware Regression IoU loss function is proposed to dynamically balance localization and shape-consistency constraints for small targets. Experimental evaluation on the VisDrone2019 benchmark demonstrates that YOLO-SWIFT achieves 38.7% mAP50 and 22.5% mAP50:95. The proposed framework provides an effective solution for intelligent aerial sensing and offers potential applications in electromagnetic imaging, remote sensing interpretation, and autonomous surveillance systems.

INDEX TERMS small object detection, wavelet transform, UAV imagery, remote sensing

I. INTRODUCTION

FLIGHT control systems has been extensively applied in diverse fields such as the identification of wild medicinal plants [1], structural damage detection in buildings [2], and emergency rescue operations [3]. The rapid development of deep learning has accelerated the convergence of UAV aerial imagery and computer vision technology, establishing a significant research trend that now extends to the high-precision defect detection and surface analysis of advanced materials. This integration enables the automated identification of structural anomalies in large-scale production and outdoor industrial significantly enhancing materials the efficiency of quality control systems. This capability enables UAVs to precisely locate and track various targets, significantly expanding their scope of application across numerous domains.

Two-stage detectors achieve high detection accuracy by focusing on potential target regions through region proposal mechanisms. However, this multi-stage cascading paradigm introduces significant inference latency, failing to meet the real-time requirements of dynamic UAV inspection scenarios. Furthermore, in ground scenarios characterized by dense and overlapping objects, generating a massive number of highly overlapping candidate boxes results in unnecessary computational resource consumption. Conversely, one-stage object detectors, leverage end-to-end architectures to achieve efficient inference. Currently, the YOLO series has become the preferred choice among object detection algorithms due to its rapid inference speed and compact model size.

While the aforementioned generic object detection methods demonstrate robust performance in natural scenes, they often fail to yield satisfactory results. For instance, in traffic scenarios, targets such as vehicles and pedestrians frequently appear at extremely small scales; shallow networks struggle to capture their finegrained features, while deep networks suffer from the loss of spatial information regarding small objects due to downsampling operations, thereby degrading detection accuracy. Modifying existing object detection frameworks has become a prevalent strategy.

Lyu et al. proposed an enhanced RT-DETR architecture specifically designed to address small object feature degradation and spatial-semantic misalignment in UAV imagery [4]. This method introduces a Channel-Aware Sensing Module that replaces traditional backbone modules with depth-

wise separable convolutions and a Convolutional Additive Token Mixer (CATM), effectively enhancing discriminative features for small objects while suppressing background noise. Addressing detection challenges in complex environments, Zhu et al. utilized deformable attention mechanisms as an auxiliary mechanism to accelerate the focusing and localization of suspicious targets under complex conditions [5]. Tang et al. proposed MS-Faster R-CNN to improve object detection and aerial tracking in UAV imagery, focusing on resolving drastic scale variations caused by changes in UAV flight altitude [6]. This method features a designed Multi-Stream CNN backbone comprising three parallel feature extraction streams, where each stream employs different kernel sizes to simultaneously capture features at varying scales and levels of detail.

To address the suboptimal performance of existing detectors in detecting small objects and handling scale variations, Wang et al. designed the DMFF neck network, which incorporates a dedicated small object detection head, utilizes a DMC module, and employs a DSSFF module to construct a scale sequence space for scale adaptation [7]. Targeting the spatial heterogeneity of sparse yet locally dense target distributions in UAV imagery, Chen et al. proposed LUD-YOLO to achieve an optimal balance between model lightweighting and detection performance [8]. Based on YOLOv8, this method introduces a novel multi-scale feature fusion pattern that addresses feature propagation degradation through an improved feature pyramid, while incorporating a Dynamic Sparse Attention mechanism within the detection module.

To address the challenges inherent in UAV aerial imagery and high-speed surface inspection— specifically, extremely small object scales, complex backgrounds, and feature loss during downsampling—this paper proposes YOLO-SWIFT, a small-object perception network improved upon YOLOv11n. This architecture effectively identifies both aerial targets and microscopic defects in technical materials, overcoming the limitations of traditional vision systems in densely distributed and variable environments. The main contributions are summarized as follows:

Development of the Position-Aware Coordinate Downsampling (PACD) Module: By incorporating a coordinate attention mechanism into the downsampling process, this module embeds essential spatial positional information into channel features. Combined with a skip connection structure to preserve the continuity of feature propagation, PACD effectively addresses the issue of positional information degradation for tiny objects, which commonly occurs during feature dimensionality reduction in conventional convolutions.

Introduction of the High-Resolution Feature Aggregation Network (HRFA-Net): An ultra-fine detection branch is appended to the neck network to construct cross-level feature fusion pathways. This architecture enables the comprehensive integration of deep abstract semantic representations with shallow high-resolution texture features, substantially enhancing the model's perception capability for extremely small objects and effectively reducing the incidence of missed detections.

Construction of the Wavelet Bottleneck Enhancement Module (WBE-C3): Employing cascaded wavelet decomposition to restructure the bottleneck layers of the backbone network, this module significantly expands the network's effective receptive field for capturing global contextual information. Concurrently, a novel High-Frequency Retention (HFR) pathway injects high-frequency texture details into the deep network, overcoming the limitation of traditional large convolution kernels, which tend to sacrifice edge details of small objects when expanding the receptive field.

Design of the Adaptive Scale-aware Regression IoU (ASRIoU) Loss Function: A dynamic weighting mechanism based on object scale is established to adaptively adjust the weight distribution between scale consistency loss and center point localization loss. This effectively suppresses severe gradient fluctuations caused by minor positional deviations of small objects, thereby improving the model's localization accuracy and convergence stability in complex scenarios.

II. RELATED WORK

A. OBJECT DETECTION IN UAV IMAGERY

Object detection in UAV aerial imagery has emerged as a critical research area due to its wide-ranging applications in surveillance, traffic monitoring, disaster response, and environmental monitoring [9]. Unlike conventional object detection scenarios, UAV imagery presents unique challenges. The VisDrone dataset [10] has become a benchmark for evaluating detection algorithms in this domain, providing diverse scenarios captured across 14 different cities with significant variations in object density and scale.

Traditional approaches to UAV object detection have focused on adapting existing detection frameworks through multi-scale feature fusion [11], attention mechanisms [12], and specialized loss functions. Liu et al. though their computational requirements limit deployment on resource-constrained UAV platforms [13]. Recent works have explored lightweight architectures specifically designed for edge deployment [14], balancing detection accuracy with computational efficiency. Attention mechanisms have become fundamental components in modern deep learning architectures. Spatial attention mechanisms complement channel attention by highlighting informative spatial regions [15].

More recently, WTConv [16] proposed using cascaded wavelet decomposition to expand the effective receptive field without increasing parameter count, achieving a balance between global context capture and local detail preservation. These developments motivate our design of the WBE-C3 module, which extends wavelet-based feature enhancement with explicit high-frequency retention pathways.

B. LOSS FUNCTIONS FOR SMALL OBJECT DETECTION

These loss functions exhibit instability when applied to small objects, where minor positional deviations cause disproportionate IoU fluctuations [4]. Scale-aware loss functions have been proposed to address this issue by dynamically adjusting loss weights based on object size [17]. The ASRIoU loss adopted in this work extends this concept by decoupling scale consistency and localization accuracy, enabling more stable gradient flow during small object training.

III. METHODOLOGY

To address the issues of low detection accuracy caused by the susceptibility of tiny object features to loss and complex backgrounds in aerial imagery, this paper proposes improvements using YOLOv11n as the baseline. The overall network architecture of YOLO-SWIFT is illustrated in Figure 1.

A. POSITION-AWARE COORDINATE DOWNSAMPLING (PACD) MODULE

UAV aerial imagery contains a vast number of small objects densely distributed within complex backgrounds. The traditional convolution modules in YOLOv11 exhibit limited capability in processing features of small objects, frequently resulting in the loss of critical feature information. To address this challenge, this paper designs the PACD (Position-Aware Coordinate Downsampling) module, the architecture of which is illustrated in Figure 2.

The skip connection path is implemented via an AvgPool2d(2×2) operation, which matches the downsampling rate of the main branch, ensuring that the feature maps of both branches possess consistent dimensions prior to element-wise addition.

B. HIGH-RESOLUTION FEATURE AGGREGATION NETWORK (HRFA-NET)

When applied to UAV aerial imagery, the baseline YOLOv11 model faces severe challenges regarding missed detections of small objects. This is primarily attributed to the inherent low resolution and limited feature information of small objects; within the standard PANet structure [18], their critical information is prone to dilution or neglect during transmission through deeper network layers. The model's default detection heads predominantly focus on medium-to-high-level feature maps, resulting in the underutilization of shallow, high-resolution features and consequently suboptimal detection performance. To address this issue, this paper proposes an improvement to the Neck component of YOLOv11 by designing the High-Resolution Feature Aggregation Network (HRFA-Net). While the PACD module in the backbone focuses on preserving spatial fidelity during initial feature downsampling, HRFA-Net provides a complementary structural solution at the neck level. Its core philosophy involves the introduction of an ultra-fine detection branch to prevent the dilution of these preserved spatial features as they propagate into deeper semantic layers. Its core philosophy involves the introduction of an ultra-fine detection branch. In the top-down pathway, medium-level feature maps, which possess strong semantic information, are restored to a higher-resolution scale via upsampling. The fused features undergo processing by a deep integration module to generate new high-resolution feature maps, which are subsequently fed into a newly added detection head dedicated to ultra-small targets. In this study, following the VisDrone dataset benchmarks, standard "small targets" are defined as those with an area of less than 32×32 pixels. We further define "ultra-tiny objects" as a specific subset involving targets smaller than 16×16 pixels (often covering fewer than 0.01% of the total image area), which typically suffer from severe feature disappearance in standard P3-P5 feature pyramids. Simultaneously, to maintain the structural integrity of the PANet, a downsampling module is correspondingly added to the subsequent layers. This enables the acquisition of feature map information with dimensions of 160×160, thereby further extracting additional shallow features pertinent to small objects. The HRFA-Net architecture effectively synergizes high-resolution shallow features with high-level semantic features, significantly enhancing the model's perception and localization capabilities for small objects and effectively mitigating the phenomenon of missed detections.

C. WAVELET BOTTLENECK ENHANCEMENT MODULE (WBE-C3)

In the YOLOv11 architecture, the C3K2 module serves as a critical component for feature extraction. Its design employs a split-fusion strategy: the input features are divided into two paths—one undergoes standard convolution, while the other is processed through multiple Bottleneck structures for deep feature extraction. By introducing variable kernel sizes and channel separation techniques, the C3K2 module aims to expand the sensing range, enabling the capture of broader contextual information to enhance feature extraction efficiency in complex scenarios.

Despite the efficiency of the C3K2 design, its core bottleneck unit still relies on traditional 3×3 convolutions.

This design harbors a fundamental limitation: the receptive field of shallow feature maps grows slowly, restricting their semantic representation capability to capturing only local information [19]. Attempts to expand the receptive field by simply increasing the convolution kernel size immediately face a quadratic surge in parameter volume. This not only risks over-parameterization but also leads to rapid performance saturation before a global receptive field is achieved.

To address this challenge, we propose a novel receptive field enhancement module, named WBE-C3, as illustrated in Figure 3. We redesigned all C3K2 modules in the YOLOv11 backbone, replacing the original 3×3 convolutions in their bottleneck units with our novel Wavelet Receptive Field Convolution (WRFConv). Inspired by WTConv [20], WR-

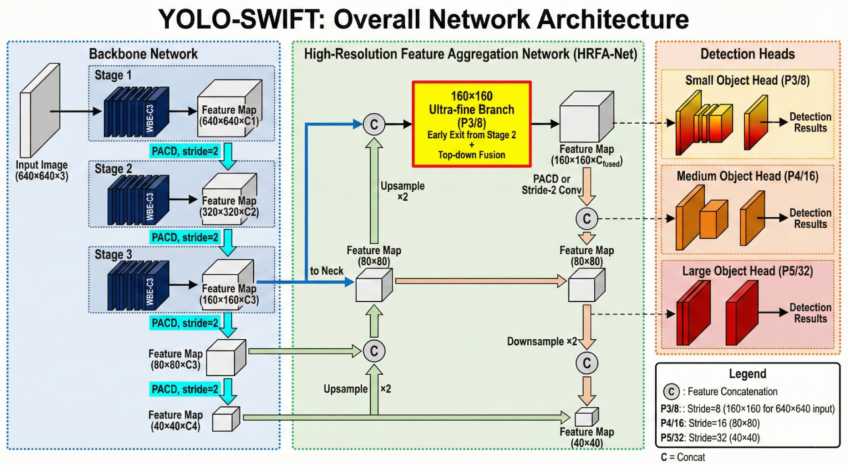

FIGURE 1. The overall network architecture of YOLO-SWIFT

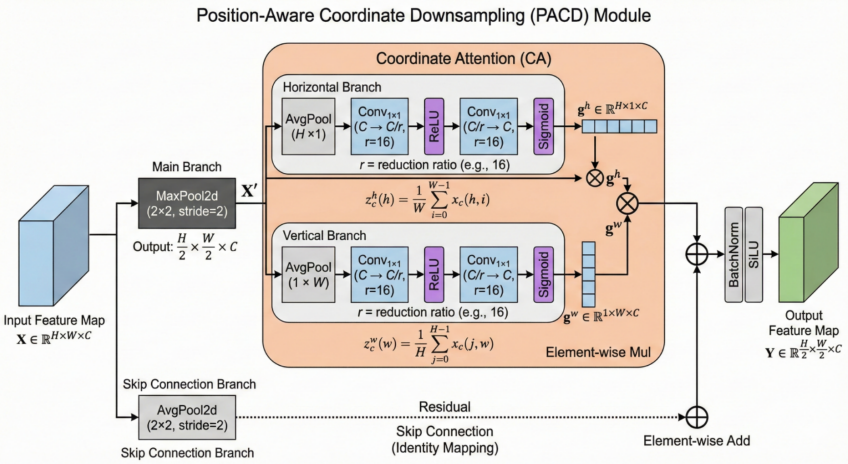

FIGURE 2. Schematic diagram of the PACD module architecture. The module comprises MaxPool2d for downsampling, Coordinate Attention for spatial encoding, and a skip connection path with AvgPool2d for feature preservation

FConv splits the input into multi-frequency sub-bands via cascaded wavelet decomposition and performs operations using fixed 3×3 convolution kernels on the progressively downsampled low-frequency (LL) components. This mechanism enables the effective receptive field of a single layer to expand exponentially with the number of decomposition levels. It systematically enhances the backbone network's receptive field range, allowing effective capture of large-scale object and shape priors at various stages, thereby mitigating detection difficulties caused by a lack of contextual information in small kernels. In the context of the VisDrone dataset, which spans 14 different cities, this wavelet-based approach is particularly crucial. The multi-level decomposition allows the network to adaptively filter out varied urban background noise (high-frequency interference) while using the expanded receptive field to capture consistent structural priors of objects across different city-level lighting and architectural environments. This ensures that the feature extraction remains robust despite the high variance in scene characteristics inherent in diverse urban locales.

While WTConv enhances shape bias by passing only low-frequency components during recursion, this approach poses a risk of losing critical high-frequency details inherent to small objects. Therefore, our WRFConv introduces a High-Frequency Retention Path (HFR). In each decomposition iteration, the high-frequency components ($Y_H^{(i)}$), after convolution, are aggregated via a 1×1 convolution (HFR module) and added back to the low-frequency component ($Y_{LL}^{(i)}$). This design ensures that high-frequency detail information is injected into deeper levels of the decomposition, achieving a unification of large receptive fields and detail preservation.

The core mathematical formulation of WRFConv is described as follows. First, cascaded decomposition is performed via Wavelet Transform (WT), where $X_{LL}^{(0)}$ represents the initial input:

$$X_{LL}^{(i)}, X_H^{(i)} = \mathrm{WT}\left(X_{LL}^{(i-1)}\right) \tag{1}$$

Next, convolutions are performed on the separated frequency components in the wavelet domain:

$$Y_{LL}^{(i)}, Y_H^{(i)} = \mathrm{Conv}\left(W^{(i)}, \left(X_{LL}^{(i)}, X_H^{(i)}\right)\right) \tag{2}$$

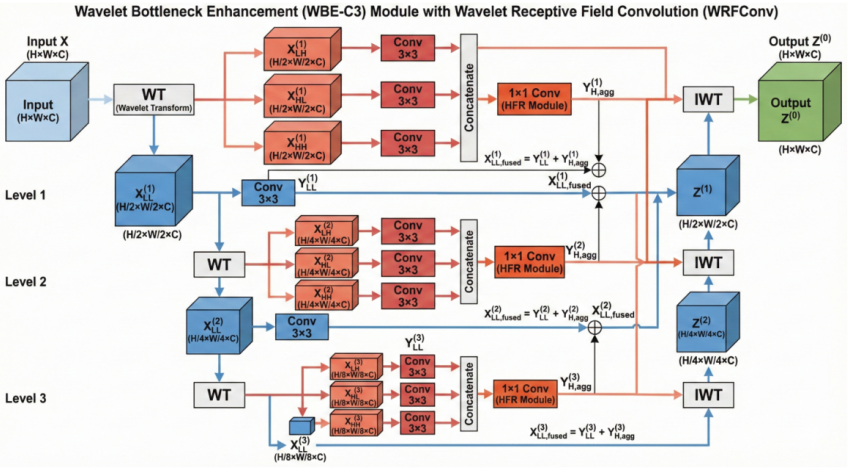

FIGURE 3. Architecture of the WBE-C3 module with Wavelet Receptive Field Convolution (WRFConv). The module employs cascaded wavelet decomposition with a High-Frequency Retention (HFR) path to maintain textural details while expanding the receptive field

Subsequently, the proposed High-Frequency Retention (HFR) operation is executed:

$$Y_{H,agg}^{(i)} = \mathrm{Conv}_{\mathrm{HFR}}^{(i)}\left(Y_H^{(i)}\right) \tag{3}$$

The aggregated high-frequency information is then fused back into the low-frequency channel:

$$X_{LL,fused}^{(i)} = Y_{LL}^{(i)} + Y_{H,agg}^{(i)} \tag{4}$$

Finally, the feature map is recursively reconstructed using the Inverse Wavelet Transform (IWT) and by aggregating the output $Z^{(i+1)}$ from the deeper layer $(i+1)$ (where $Z^{(L+1)}$ is initialized to 0):

$$Z^{(i)} = \mathrm{IWT}\left(X_{LL,fused}^{(i)} + Z^{(i+1)}, Y_H^{(i)}\right) \tag{5}$$

The final output $Z^{(0)}$ represents the result of the WRFConv module, which achieves a parameter-efficient global receptive field while preserving high-frequency details within a single layer.

D. ADAPTIVE SCALE-AWARE REGRESSION IOU (ASRIOU) LOSS FUNCTION

CIoU constructs a more comprehensive metric. Building upon IoU, it incorporates three critical geometric factors. The penalty term αv is calculated as follows:

$$\begin{aligned} v &= \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2, \\ \alpha &= \frac{v}{(1-\mathrm{IoU})+v} \end{aligned} \tag{6}$$

Despite providing a more refined metric than standard IoU, CIoU exhibits severe limitations when applied to small object detection tasks in UAV aerial imagery. For small objects occupying only a few pixels, even a positional deviation of one or two pixels in the predicted box can trigger drastic and disproportionate fluctu ations in the IoU value. This high sensitivity to minute positional errors generates unstable gradients, which not only interfere with the model's normal convergence but also amplify the negative impact of inevitable labeling noise within the dataset.

ASRIoU is specifically designed to enhance the regression stability and accuracy for small-scale objects. The core mechanism of ASRIoU is its dynamic weighting system, which adaptively adjusts the loss calculation method based on the absolute scale of the object. It decouples the regression task into two primary components:

Scale Loss ($\mathcal{L}_{SC}$): An IoU-based loss used to measure the consistency of size and shape:

$$\mathcal{L}_{SC} = 1 - \mathrm{IoU} + \alpha v \tag{7}$$

Localization Loss ($\mathcal{L}_{LC}$): A distance-based loss used to measure the localization accuracy of the center point:

$$\mathcal{L}_{LC} = \frac{\rho^2(b_p, b_{gt})}{c^2} \tag{8}$$

The key innovation of ASRIoU lies in the fact that it does not simply sum these two loss components; instead, it introduces a scale-aware mechanism to dynamically allocate their weights. To accurately assess the true scale of the target, the scaling factor R_{oc} between the original image and the current feature map is first calculated:

$$R_{oc} = \frac{\omega_0 \times h_0}{\omega_c \times h_c} \tag{9}$$

where ω_0, h_0 are the width and height of the original image, and ω_c, h_c are those of the current feature map. Next, this scaling factor is utilized to calculate the bounding box impact coefficient β_B, which quantifies the influence of object scale on the loss calculation:

$$\beta_B = \min\left(\frac{B_{gt}}{B_{gt}^{\max}} \times R_{oc} \times \delta, \delta\right) \tag{10}$$

where B_{gt} is the area of the ground-truth box, $B_{gt}^{\max} = 81$ is the preset maximum object size (defining the threshold

for small objects), and δ is a tunable dynamic coefficient used to constrain the range of the impact coefficient.

Finally, the ASRIoU loss ($\mathcal{L}_{ASR}$) dynamically calculates the weight factors $\beta_{\mathcal{L}_{SC}}$ and $\beta_{\mathcal{L}_{LC}}$ for $\mathcal{L}_{SC}$ and $\mathcal{L}_{LC}$ via β_B and δ:

$$\beta_{\mathcal{L}_{SC}} = 1 - \delta + \beta_B, \quad \beta_{\mathcal{L}_{LC}} = 1 + \delta - \beta_B \tag{11}$$

$$\mathcal{L}_{ASR} = \beta_{\mathcal{L}_{SC}} \times \mathcal{L}_{SC} + \beta_{\mathcal{L}_{LC}} \times \mathcal{L}_{LC} \tag{12}$$

Analysis of the Mechanism: When the target bounding box area is smaller than 81, the value of β_B approaches 0. Consequently, $\beta_{\mathcal{L}_{SC}}$ approaches $1 - \delta$, while $\beta_{\mathcal{L}_{LC}}$ approaches $1 + \delta$. This means the model reduces its focus on the unstable IoU loss ($\mathcal{L}_{SC}$) for small objects and shifts its emphasis toward optimizing the accuracy of center point localization ($\mathcal{L}_{LC}$). Conversely, when the target bounding box area is larger than 81, $\beta_B = \delta$, resulting in $\beta_{\mathcal{L}_{SC}} = \beta_{\mathcal{L}_{LC}} = 1$. In this case, $\mathcal{L}_{ASR}$ degenerates into the standard CIoU loss, attending to both scale and localization simultaneously. This dynamic adjustment mechanism effectively mitigates the problem of gradient instability caused by violent IoU fluctuations in small object detection, while simultaneously suppressing the negative effects of label noise, thereby significantly improving the detection robustness and localization accuracy of the model on UAV aerial imagery.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL DATASET

Significant characteristics of this dataset include a wide dynamic range of image resolutions (spanning from 480×360 to 2000×1500 pixels) and a high proportion of small object instances, which account for approximately 60% of the data. Furthermore, severe object occlusion and dense distribution are prevalent. These factors render VisDrone2019 a highly challenging benchmark that is exceptionally well-aligned with the objectives of this research.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{13}$$

Depending on the Intersection over Union (IoU) threshold employed, two specific metrics are commonly used: mAP@0.5 and mAP@0.5 : 0.95.

B. ABLATION STUDIES

Compared with the original YOLOv11n model, the improved YOLO-SWIFT incorporates four key enhancements: the introduction of the Position-Aware Coordinate Downsampling module PACD (A), the design of the High-Resolution Feature Aggregation Network HRFA-Net (B), the optimization of the C3K2 module into WBE-C3 (C), and the adoption of the ASRIoU loss function (D). The experiments were performed by progressively integrating each improvement component into the baseline model to observe their impact on performance. The results of the ablation studies are presented in Table 3.

TABLE 3. Ablation study results based on YOLOv11n

Model	A	B	C	D	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)
YOLOv11n					43.4	33.6	31.5	18.3
YOLOv11n	✓				44.9	34.5	33.6	19.2
YOLOv11n	✓	✓			46.8	36.8	36.9	21.5
YOLOv11n	✓	✓	✓		48.1	37.9	38.2	22.3
YOLOv11n	✓	✓	✓	✓	48.6	38.4	38.7	22.5

An analysis of the experimental results in Table 3 and Figure 4 reveals the following:

Compared to the benchmark YOLOv11n model, the model incorporating the PACD module (A) achieved improvements of 2.3% and 1.4% in mAP50 and mAP50:95, respectively. This indicates that by integrating the coordinate attention mechanism into the downsampling process, the PACD module successfully resolves the issue of positional information loss inherent in traditional convolutions during feature map reduction, enabling the network to more acutely perceive the spatial coordinates of small objects within complex backgrounds.

Furthermore, upon the inclusion of the High-Resolution Feature Aggregation Network (HRFA-Net) (B) on top of the PACD, mAP50 and mAP50:95 saw substantial increases of 3.3% and 2.3%, respectively. This demonstrates that by introducing an ultra-fine detection branch, HRFA-Net fully fuses high-resolution shallow texture information with high-level semantic features, effectively mitigating the issues of high missed detection rates and the tendency to overlook ultra-small targets in UAV aerial scenarios. This suggests that the WBE-C3 module, by leveraging wavelet convolution technology, locks onto the high-frequency texture edge details of small objects via the High-Frequency Retention (HFR) path while expanding the receptive field, thereby further reinforcing feature discriminability.

Finally, with the network structure remaining unchanged, replacing the bounding box regression loss with the Adaptive Scale-aware Regression IoU loss (ASRIoU) (Model D) further elevated mAP50 and mAP50:95 to 38.7% and 22.5%. This proves that ASRIoU, through its scale-aware dynamic weighting mechanism, effectively suppresses the excessive sensitivity of small object IoU to positional deviations, enhancing the stability and precision of bounding box regression.

To further demonstrate the training stability advantages of ASRIoU, Figure 5 presents the training loss convergence curves comparing CIoU and ASRIoU loss functions. As illustrated, the proposed ASRIoU loss achieves convergence at epoch 85, significantly earlier than CIoU which converges at epoch 120. Moreover, ASRIoU attains a 26.1% lower final loss value (0.148 vs. 0.201), with notably reduced variance throughout training. These results validate that the scale-aware dynamic weighting mechanism effectively mitigates

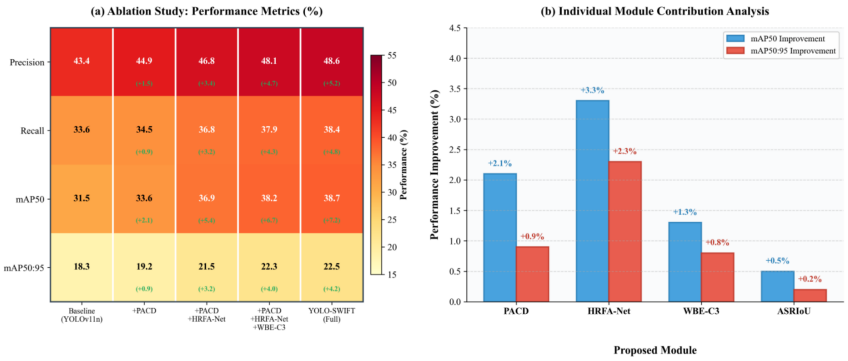

FIGURE 4. Visualization of ablation study results. (a) Performance metrics heatmap showing progressive improvements with each module addition; (b) Individual module contribution analysis

gradient instability caused by IoU fluctuations in small object detection.

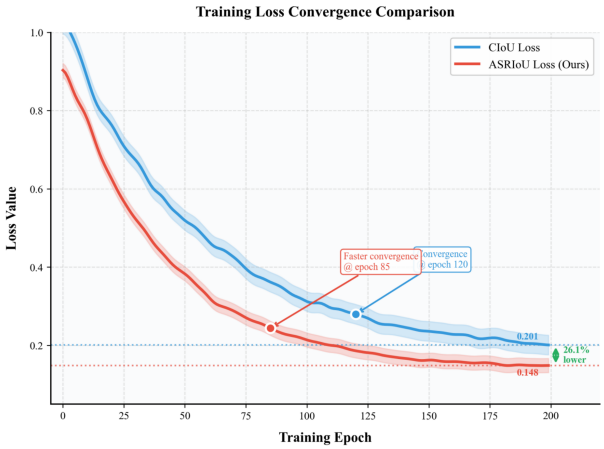

FIGURE 5. Training loss convergence comparison between CIoU loss and ASRIoU loss

In summary, the four improvements—PACD, HRFA-Net, WBE-C3, and ASRIoU—all made significant contributions to detection performance.

C. CATEGORY-WISE PERFORMANCE ANALYSIS

Figure 6 presents a detailed comparison of the Average Precision (AP) between YOLOv11n and YOLO-SWIFT across ten object categories.

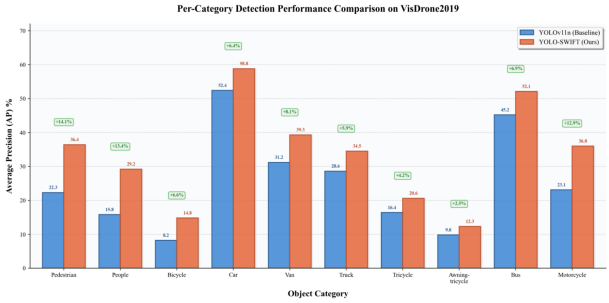

FIGURE 6. Per-category AP comparison between YOLOv11n and YOLO-SWIFT

Specifically, within the small object categories, YOLO-SWIFT significantly outperforms YOLOv11n in detection accuracy for pedestrians, people, bicycles, tricycles, and awning-tricycles, with improvements of 14.1%, 13.4%, 6.6%, 4.2%, and 2.5%, respectively. This validates the algorithm's remarkable robustness and balance in handling multi-scale objects.

D. COMPARATIVE EXPERIMENTS

To verify the effectiveness and superiority of the proposed YOLO-SWIFT algorithm and to comprehensively analyze and evaluate its performance in object detection tasks from an aerial perspective, representative classic algorithms, YOLO series algorithms, and recent improved algorithms in the object detection field were selected for comparative experiments. All algorithms were trained and tested on the VisDrone2019 dataset. The results of these comparative experiments are presented in Table 4.

TABLE 4. Comparative experiments on the VisDrone2019 dataset

Model	Precision (%)	Recall (%)	mAP50 (%)	mAP50:95 (%)	Params (M)
SSD (Liu et al. 2016)	21.1	35.8	24.0	11.9	12.3
Faster R-CNN (Ren et al. 2015)	34.6	33.8	33.3	20.6	41.0
QueryDet (Yang et al., 2022)	41.4	33.4	31.6	17.4	18.9
YOLOv5s (Jocher et al. 2022)	47.5	36.5	37.8	21.3	9.1
YOLOv7-tiny (Wang et al., 2023)	47.0	36.2	35.3	19.6	6.02
YOLOv8n (Jocher et al. 2023)	43.65	32.9	32.8	19.1	3.01
YOLOv8s (Jocher et al. 2023)	47.9	37.5	38.0	21.6	11.2
YOLOv10n (Wang et al., 2024)	44.2	34.2	34.2	19.8	2.26
YOLOv11n	43.4	33.6	31.5	18.3	2.58
TPH-YOLOv5 (Zhu et al., 2021)	–	–	36.3	21.8	3.5
CEASC (Chen et al., 2023)	–	–	37.0	21.4	4.4
YOLO-SWIFT (Ours)	48.3	38.4	38.7	22.5	2.55

As indicated by the experimental data in Table 4, YOLO-SWIFT achieves an mAP50 of 38.7%. With a parameter count of only 2.55M, its overall performance outperforms most comparison methods.

Comparison with Classic Detectors: Compared with classic detectors such as SSD and Faster R-CNN, YOLO-SWIFT demonstrates significant advantages across various accuracy evaluation metrics, surpassing them in mAP50 by 14.7% and 5.4%, respectively.

Comparison with Specialized Detectors: YOLO-SWIFT achieves further improvements in detection accuracy with fewer parameters, increasing mAP50 by 7.1%. Comparison with YOLO Series (v5/v7/v8): Compared to a series of YOLO models, the proposed YOLO-SWIFT demonstrates significant advantages. Relative to YOLOv5s and YOLOv8s, YOLO-SWIFT's mAP50 is higher by 0.9 and 0.7 percentage points, respectively, while its parameter count is only about 28% of YOLOv5s and 23% of YOLOv8s.

Comparison with Lightweight Models: Furthermore, when compared with the lightweight YOLOv8n, YOLOv10n, and YOLOv11n, although the parameter scales are of the same order of magnitude, YOLO-SWIFT outperforms them in mAP50 by 5.9, 4.5, and 7.2 percentage points, respectively. This indicates that its accuracy advantage is even more pronounced under lightweight settings. Furthermore, the performance gains observed in YOLO-SWIFT suggest that the proposed architectural enhancements—particularly the ASRIoU loss and PACD module—are not strictly dependent on the YOLOv11 baseline. Due to their modular design targeting fundamental small-object detection bottlenecks (e.g., gradient instability and spatial information loss), these components possess the potential to yield similar accuracy improvements when integrated into other YOLO series versions, such as YOLOv8 or YOLOv10.

Comparison with SOTA Small Object Models: For TPH-YOLOv5 and CEASC, which target similar small object scenarios, YOLO-SWIFT also achieved superior results. With slightly lower or comparable parameter counts, its mAP50 is higher by 2.4 and 1.7 percentage points, respectively.

Synthesizing the comparisons across all groups, it can be concluded that YOLO-SWIFT balances high detection accuracy with low model complexity in UAV aerial small object detection tasks, possessing significant engineering application value.

E. VISUALIZATION RESULTS ANALYSIS

To visually compare the detection performance of the proposed YOLO-SWIFT algorithm against the baseline model YOLOv11n in UAV aerial scenarios, four representative complex scenarios were selected from the VisDrone2019 dataset as test samples. These include: scenarios with highly dense distributions of small objects, scenarios with drastic illumination changes, low-light nighttime scenarios, and high-altitude long-distance aerial scenarios. The visualization results are presented in Figure 7.

High-Density Scenarios: The first row illustrates a high-density object environment. The scene features crowded people, mutual occlusion, and significant overlap of small objects. The baseline algorithm, YOLOv11n, suffers from frequent false and missed detections. In contrast, YOLO-SWIFT significantly reduces these errors and is capable of identifying objects that YOLOv11n failed to detect.

Complex Lighting Scenarios: In scenarios with complex lighting conditions (second row), leading to false detections. YOLO-SWIFT demonstrates stronger adaptability, effectively suppressing background interference and focusing more accurately on the targets themselves.

Nighttime Scenarios: In the nighttime detection shown in the third row, due to insufficient lighting, blurred object features, and indistinguishable details, YOLOv11n exhibits obvious missed detections and misclassifications.

High-Altitude Scenarios: In high-altitude aerial scenarios, objects contain fewer pixels, features are less distinct, and background information is rich. The fourth row shows that the baseline model faces issues with failing to detect some small objects or inaccurate category judgments. Conversely, YOLO-SWIFT is able to more accurately identify specific categories such as vehicles and pedestrians, reducing missed and false detections of distant small objects.

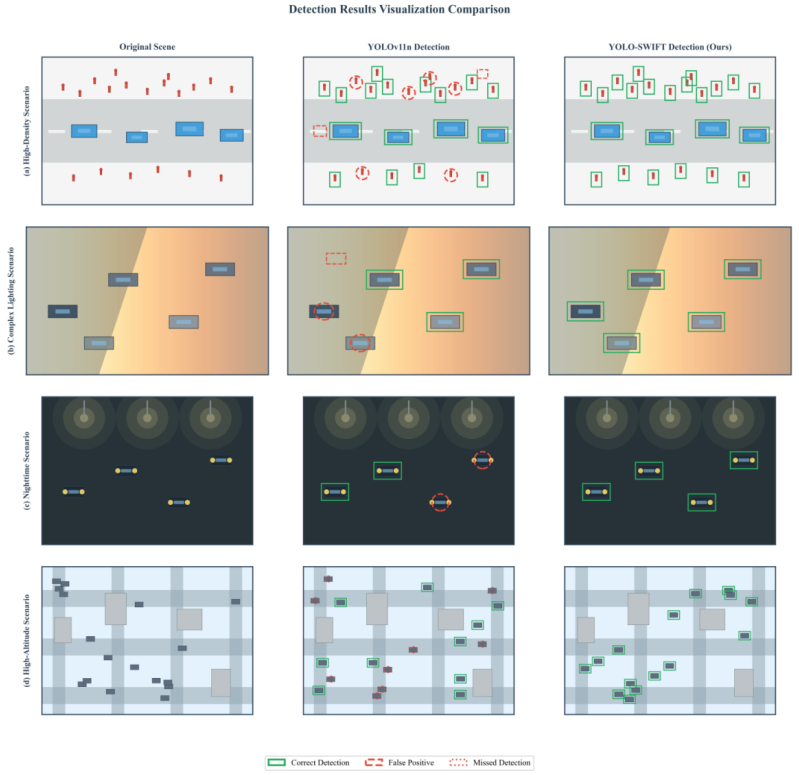

FIGURE 7. Visualization comparison of detection results. Left column: original images; middle column: YOLOv11n results; right column: YOLO-SWIFT results

V. CONCLUSION

YOLO-SWIFT effectively resolves challenges inherent to UAV aerial imagery and high-speed surface inspection, including minute object dimensions, severe background noise interference, and the susceptibility of features to loss during downsampling, thereby significantly enhancing detection performance for both aerial targets and microscopic material defects.This unified perception framework ensures highprecision identification across complex environments, ranging from expansive UAV perspectives to the intricate textures of industrial manufacturing [21].

First, the PACD module was designed. By integrating a coordinate attention mechanism with a skip connection structure, it encodes spatial positional information into channel features during the downsampling process. Second, the WBE-C3 module was constructed. Utilizing cascaded wavelet decomposition and a High-Frequency Retention (HFR) path, it locks onto the texture details of small objects while substantially expanding the effective receptive field, overcoming the inherent contradiction between parameter volume and receptive field size in traditional convolutions [22].

Third, the HRFA-Net structure was proposed. By adding an ultra-fine detection branch and establishing cross-level fusion channels, it effectively leverages shallow texture information, significantly enhancing the perception capability for ultra-tiny objects.

Finally, the ASRIoU loss function was introduced. Through a scale-aware mechanism that dynamically adjusts loss weights, it suppresses gradient fluctuations caused by positional deviations of tiny objects.

AUTHOR CONTRIBUTIONS

Conceptualization –Yanchen Li and Kai Shao; methodology – Yanchen Li and Kai Shao; investigation – Yanchen Li and Kai Shao; writing-original draft preparation – Yanchen Li and Kai Shao. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, “Review on convolutional neural networks (CNN) in vegetation remote sensing,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, pp. 24-49, 2021, doi: 10.1016/j.isprsjprs.2020.12.010.

[2] B. F. Spencer, V. Hoskere, and Y. Narazaki, “Advances in computer vision-based civil infrastructure inspection and monitoring,” Engineering, vol. 5, no. 2, pp. 199-222, 2019, doi: 10.1016/j.eng.2018.11.030.

[3] B. Mishra, D. Garg, P. Narang, and V. Mishra, “Drone-surveillance for search and rescue in natural disaster,” Computer Communications, vol. 156, pp. 1-10, 2020, doi: 10.1016/j.comcom.2020.03.012.

[4] C. Lyu, W. Zhang, H. Huang, et al., "RTMDet: An empirical study of designing real-time object detectors," arXiv preprint arXiv:2212.07784, 2022.
[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," In International Conference on Learning Representations, 2020.
[6] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "GhostNetV2: Enhance cheap operation with long-range attention," Advances in Neural Information Processing Systems, vol. 35, pp. 9969-9982, 2023, doi: 10.52202/068431-0724.
[7] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, K. Han, and Y. Wang, "Gold-YOLO: Efficient object detector via gather-and-distribute mechanism," Advances in Neural Information Processing Systems, vol. 36, pp. 51094-51106, 2023, doi: 10.52202/075280-2224.
[8] Z. Chen, C. Yang, Q. Li, F. Zhao, Z. J. Zha, and F. Wu, "Disentangle your dense object detector," In Proceedings of the 31st ACM International Conference on Multimedia, pp. 4939-4950, 2023, doi: 10.1145/3474085.3475351.
[9] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 2778-2788, 2021, doi: 10.1109/ICCVW54120.2021.00312.
[10] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.
[11] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117-2125, 2017, doi: 10.1109/ CVPR.2017.106.
[12] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," In Proceedings of the European Conference on Computer Vision, pp. 3-19, 2018, doi: 10.1007/978-3-030-01234-2_1.
[13] Z. Liu, Y. Lin, Y. Cao, et al., "Swin transformer: Hierarchical vision transformer using shifted windows," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022, 2021, doi: 10.1109/ ICCV48922.2021.00986.
[14] A. Howard, M. Sandler, G. Chu, et al., "Searching for MobileNetV3," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314-1324, 2019, doi: 10.1109/ICCV.2019.00140.
[15] S. E. Finder, R. Amoyal, E. Treister, and O. Freifeld, "Wavelet convolutions for large receptive fields," In European Conference on Computer Vision, pp. 351-367. Springer, 2024, doi: 10.1007/978-3-031-72949-2_21.
[16] X. Li, W. Wang, L. Wu, et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," Advances in Neural Information Processing Systems, vol. 33, pp. 21002-21012, 2020.
[17] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759-8768, 2018, doi: 10.1109/CVPR.2018.00913.
[18] Han Y , Wang C , Luo H ,et al.LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design[J].SCIENTIFIC REPORTS, vol. 15, no. 1, 2025, doi: 10.1038/ s41598-025-07021-6.
[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141, 2018, doi: 10.1109/CVPR.2018.00745.
[20] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 773-782, 2019, doi: 10.1109/ CVPRW.2018.00121.
[21] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464-7475, 2023, doi: 10.1109/CVPR52729.2023.00721.
[22] A. Wang, H. Chen, L. Liu, et al., "YOLOv10: Real-time end-to-end object detection," arXiv preprint arXiv:2405.14458, 2024.