

Article

FFR-YOLO: A Frequency-Guided Fusion Reconstruction Network for Small-Object Detection in Remote Sensing Images

Pengfei Zhang , Jianqiang Zhang, Jian Liu, Xingda Li , Yiping Liu and Ling Tan 

College of Weaponry Engineering, Naval University of Engineering, Wuhan 430030, China; 0909061042@nue.edu.cn (J.Z.); 0909061051@nue.edu.cn (J.L.); m24382604@nue.edu.cn (X.L.); d25382602@nue.edu.cn (Y.L.); d25382609@nue.edu.cn (L.T.)

* Correspondence: m23181105@nue.edu.cn

Highlights

What are the main findings?

- A frequency-guided fusion reconstruction network, termed FFR-YOLO, is proposed for small-object detection in remote sensing images. Through the synergistic enhancement of three modules—FAPD, BRFM, and RDLGF—the network effectively boosts the capabilities of small-object-feature extraction, multi-scale fusion, and detector head feature recalibration.
- On the DIOR and NWPU VHR-10 datasets, FFR-YOLO achieves certain performance improvements compared to the baseline YOLOv8, delivering superior detection accuracy with only marginal increases in parameter count and computational complexity.

What are the implications of the main findings?

- The study demonstrates that the integration of the FAPD, BRFM, and RDLGF modules can significantly enhance the robustness and generalization capability of remote sensing detection models.
- The proposed framework possesses strong practical value, offering technical support for intelligent object detection tasks in scenarios, such as remote sensing monitoring, urban planning, disaster assessment, and resource exploration.

Abstract

To address the challenges of small scales, weak features, complex backgrounds, and misalignment in multi-scale fusion for small-object detection in remote sensing images, this study proposes a frequency-guided fusion reconstruction YOLO (FFR-YOLO), an improved YOLOv8 framework. The method performs joint optimization across three levels: the backbone, neck, and front end of the detector head. In the backbone, a frequency-guided anti-alias progressive downsampling module utilizes Haar wavelet decomposition to replace traditional strided convolutions and incorporates a low-frequency-guided high-frequency gating mechanism to mitigate detail loss and background noise interference during downsampling. In the neck, a bridge-guided bidirectional reconstruction fusion module (BRFM) enhances the collaborative reconstruction of multi-scale semantic and detailed information via multi-source weighted fusion and cross-path bridging interactions. At the front end of the detector head, a recalibrated dual-branch local–global fusion (RDLGF) module implements dynamic allocation and complementary fusion of dual-path features. Experiments were conducted on two datasets, DIOR and NWPU VHR-10. The results demonstrate that FFR-YOLO achieves a mAP@0.5 of 85.8% and a mAP@0.5:0.95 of 63.1% on DIOR and 93.6% and 62.1% on NWPU VHR-10. These outcomes present



Academic Editor: Zhenming Peng

Received: 23 June 2026

Revised: 20 August 2026

Accepted: 21 August 2026

Published: 24 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

improvements over the baseline YOLOv8, validating the effectiveness and practical value of the proposed method for small-object detection in remote sensing scenarios.

Keywords: frequency-guided fusion reconstruction; remote sensing object detection; YOLOv8

1. Introduction

Remote sensing imagery, due to its wide-area coverage and periodic acquisition capabilities, has become a vital data source for tasks, such as dynamic monitoring of surface morphology, resource exploration, urban planning, and disaster assessment [1,2]. With the continuous improvement of satellite sensor resolution, the demand for automatic identification and localization of specific objects from such imagery has become increasingly urgent [3,4]. However, remote sensing scenarios are characterized by vast coverage areas and highly complex background, along with significant object-scale variations. Classifying these objects often relies heavily on abrupt edge transitions, corner responses, elongated contours, and subtle yet relatively stable local differences relative to the surrounding background [5,6]. Notably, objects such as vehicles, small vessels, and ancillary facilities around buildings often occupy only dozens or even a few pixels in remote sensing images, falling into the category of typical small objects [7,8]. Such objects are generally characterized by weak features, low contrast, and insufficient information, which poses considerable difficulties for accurate detection [9,10].

Traditional object detection methods primarily rely on hand-crafted feature descriptors (e.g., HOG [11], SIFT [12]) for object representation. However, constrained by their limited feature expression capability and scene adaptability, these methods struggle to cope with multi-scale variations and background clutter interference in remote sensing images [13]. Deep learning has fundamentally transformed this situation, with convolutional neural networks (CNNs) enhancing detection robustness through automated feature learning [14]. Single-stage detection algorithms, represented by the YOLO series, have been widely applied in remote sensing object detection due to their ability to balance both detection accuracy and inference speed [15]. However, existing methods still exhibit several limitations when processing remote sensing imagery. On the one hand, conventional convolutions have limited capacity to preserve high-frequency details, and convolutional downsampling operations are prone to inducing feature aliasing. This consequently results in blurred edges, local contour degradation, and the adhesion of adjacent objects [16,17]. Conversely, detection frameworks represented by YOLOv8 utilize multi-scale fusion paths that are primarily tailored toward semantic aggregation requirements in generic natural scenes. In remote sensing images, such designs are prone to causing missed detections of dense small objects and increasing the risk of false positives in complex backgrounds [18,19].

To address these challenges, extensive research has been conducted from various perspectives. Feature pyramid networks (FPN) [20] and their derivatives (e.g., PANet [21], BiFPN [22]) enhance multi-scale representation capabilities by integrating shallow, high-resolution spatial details with deep semantic context. These architectures narrow the semantic gap between different levels through a combination of upsampling and downsampling operations. However, conventional upsampling methods are prone to inducing spatial and semantic misalignment, whereas repetitive downsampling further exacerbates the loss of fine-grained localization information [23,24].

Although some studies have introduced attention mechanisms or deformable convolutions to improve the quality of feature alignment [25,26], their effectiveness in mitigating feature misalignment and information loss remains limited because they still primarily rely

on traditional sampling operators. Furthermore, remote sensing imaging is inherently susceptible to interference from uncontrollable factors such as platform motion, atmospheric turbulence, and scene layout [27], which inevitably blurs object boundaries and exacerbates the challenge of distinguishing small targets from complex backgrounds. To overcome these intertwined obstacles several specialized feature compensation strategies have been tailored specifically for remote sensing scenarios. For instance, the Lightweight Modality Compensation Network (LMCNet) addresses missing-modality conditions through knowledge distillation to preserve discriminative capability under limited input sources [28]; the multi-scale rotation-invariant Haar-like feature integrated CNN (MSRIHL-CNN) explicitly extracts and fuses low-level texture and edge features with high-level semantic representations to enhance target discrimination in complex multi-target environments [29]. The multikernel-size feature fusion-based CNN (MKSFF-CNN) employs parallel convolution kernels of varying sizes to capture both local textural nuances and global contour structures, achieving comprehensive feature representation through optimal multi-level fusion [30]. Crucially, the key challenges of small object detection in remote sensing are manifested in two main aspects: insufficient feature representation and background confusion. The former is closely related to information loss during downsampling, while the latter stems from semantic misalignment and background interference during cross-scale fusion.

This study develops a frequency-guided fusion reconstruction object detection network, termed as FFR-YOLO, by performing joint optimization across three components within the YOLOv8 framework: the backbone, neck, and front end of the detector head. This network features three core innovative designs: frequency-guided anti-alias progressive downsampling module (FAPD), bridge-guided bidirectional reconstruction fusion module (BRFM), and recalibrated dual-branch local–global fusion module (RDLGF). Specifically, the FAPD module is deployed at the four downsampling stage nodes in the backbone, where it utilizes Haar wavelet decomposition to replace traditional strided convolutions for spatial compression. A low-frequency-guided high-frequency gating mechanism is also introduced to enhance the preservation of small object edge and contour information while suppressing invalid high-frequency responses in the background. The bridge-guided bidirectional reconstruction fusion module (BRFM) reconstructs the feature fusion paths in the neck from two dimensions: it introduces a weighted multi-source fusion mechanism during the bottom-up stage to adaptively weight the three-way information and incorporates bridging fusion units between the top-down and bottom-up paths to enhance the collaborative reconstruction capability of semantic and detailed information. The RDLGF module is deployed at the front end of the detector head to perform feature recalibration and adaptive fusion of dual-path features at the same scale, achieving dynamic allocation of the two-way features through a collaborative local and global gating mechanism. Functionally, these three improvements constitute a seamlessly connected, collaborative optimization framework. Experimental results on multiple publicly available remote sensing datasets demonstrate that FFR-YOLO can effectively boost small-object-detection performance in complex remote sensing scenarios and exhibits favorable detection results in scenes with dense objects.

The main contributions of this study are summarized as follows:

- (1) Designed the FAPD: By incorporating Haar wavelet decomposition and a low-frequency-guided high-frequency gating mechanism, this module balances the preservation of structural information and noise suppression during resolution compression, thereby mitigating the loss of small object features caused by conventional strided convolutional downsampling.
- (2) Proposed the BRFM: Through improvements at two levels—weighted multi-source fusion and cross-path bridging interactions—this module enhances the neck network's

capability to coordinate multi-scale features and improves the fusion performance of intermediate-layer features.

- (3) Constructed the RDLGF: Utilizing a spatial-channel decoupled gating and complementary fusion strategy, this module enables the network to dynamically adjust the contribution intensity of information from different branches based on the input content, thus improving the detector head's localization performance for objects with blurred contours and low contrast.

2. Related Works

2.1. Evolution of Object-Detection Methods and Insights for Remote Sensing Scenarios

Object-detection methodologies have undergone a comprehensive evolution from hand-crafted feature engineering to deep learning. Traditional schemes, such as the early Viola–Jones framework [31] and the Histogram of Oriented Gradients (HOG) [32], heavily rely on sliding windows and hand-crafted feature extraction. Although effective in predefined, constrained environments, their feature representation capability was intrinsically limited by prior knowledge, rendering them inadequate for coping with complex backgrounds and multi-scale variations inherent in remote sensing imagery. The application of CNNs significantly enhanced detection accuracy. R-CNN, proposed by Girshick et al. [33], combined selective search with CNN feature extraction, achieving a significant improvement in detection precision. Subsequently, Fast R-CNN [34] introduced the region of interest (RoI) pooling layer to enable feature sharing, while Faster R-CNN [35] substituted external proposal generation with a region proposal network to construct an end-to-end framework, further consolidating the accuracy advantages of two-stage detection architectures. Cascade R-CNN [36] enhances proposals' quality through cascaded threshold optimization, demonstrating favorable localization precision in dense remote sensing scenes with small objects. Concurrently, the single-stage detection algorithm SSD [37] predicted object classification and coordinates by presetting anchor boxes across multi-level feature maps, effectively balancing detection speed and small object recall rates. Furthermore, the densely connected mechanism proposed in DenseNet [38] enhanced feature reuse and gradient propagation; its underlying philosophy has provided valuable insights for preserving fine-grained features in subsequent detection networks. In 2016, Redmon et al. proposed YOLOv1 [39], which uniformly modeled the object-detection problem from a regression perspective. By dividing the image into an $S \times S$ grid of cells to directly output bounding box coordinates and class confidence scores, it significantly accelerated detection speed; however, its capability for dense small-object detection and its localization accuracy required further improvement. YOLOv2 [40] introduced anchor-box mechanisms and batch-normalization strategies. YOLOv3 [41] constructed the Darknet-53 backbone and integrated a feature pyramid architecture. YOLOv4 [42] adopted CSPDarknet53 and the Mish activation function to optimize feature extraction efficiency. YOLOv5 [43] enhanced engineering deployment flexibility through adaptive anchor-box calculation and hybrid data augmentation, with its modular design supporting multi-scale models adapted to varying hardware requirements. Furthermore, YOLOv6 [44] and YOLOv7 [45] focused on re-parameterization design and dynamic label assignment strategies, respectively, achieving favorable detection performance on the VisDrone2019 remote sensing dataset. YOLOv8 [46] integrates a task-aligned assigner and dynamic non-maximum suppression, demonstrating outstanding performance in dense object scenarios. YOLOv9 [47] introduces programmable gradient information to optimize feature propagation paths, combined with a generalized efficient layer aggregation network to enhance multi-scale feature interactions. Furthermore, YOLOv10 [48] implements a non-maximum

suppression (NMS)-free training strategy, utilizing a consistent dual assignment loss to suppress redundant predictive bounding boxes.

Compared to generic scenes, object detection in remote sensing imagery faces more intricate technical challenges: Object scales span multiple orders of magnitude. Spatial distribution density is highly non-uniform, and background geographic features are intensely coupled with foreground objects in terms of texture and color characteristics. Although two-stage detectors have inherent advantages in accuracy, a significant contradiction persists between their computational complexity and the real-time processing demands of remote sensing applications. Single-stage methods have emerged as the mainstream choice in this field owing to their end-to-end inference efficiency; however, the top-down propagation mechanism of the feature pyramid network [49] renders shallow-layer small object information susceptible to attenuation caused by high-level semantic interference during the fusion process. In recent years, the introduction of the transformer architecture has provided a fresh methodological perspective for object detection. Detection transformer (DETR) [50] and deformable DETR [51] realize end-to-end detection by utilizing an encoder-decoder architecture paired with a deformable attention mechanism. Real-time detection transformer (RT-DETR) [52] blends the respective advantages of CNNs and attention mechanisms, balancing both inference speed and detection accuracy in remote sensing scenarios. Wang et al. [53] provided a systematic review of the latest advancements in oriented-object detection within optical remote sensing imagery. They reviewed the evolutionary path from horizontal-bounding-box detection to oriented-bounding-box detection, thereby offering a valuable reference for understanding the unique characteristics and research trends of remote sensing object detection.

2.2. Small-Object-Feature Enhancement and Multi-Scale Fusion in Remote Sensing Imagery

The technical difficulties of small object detection in remote sensing imagery primarily stem from the combined effects of multiple factors: The limited number of pixels occupied by target results in a low signal-to-noise ratio. Downsampling operations of deep convolutional networks continuously erode detailed representations, and interference of similar textures in complex backgrounds escalates the risk of false positives. Koyun et al. [54] demonstrated that local textures in remote sensing imagery of farmland or forest areas could be misclassified as small man-made objects, highlighting the impact of background confusion on small-object-detection performance. To address the preceding challenges, the academic community has conducted systematic research along two main lines: feature enhancement and multi-scale fusion.

As a representative work in multi-scale representation learning, the feature pyramid network [49] achieves cross-level feature fusion by constructing top-down lateral connections; however, its unidirectional information flow limits sufficient interaction between shallow-layer details and deep-layer semantics. Based on this, the path aggregation network [55] adds a bottom-up enhancement path to shorten the propagation distance of shallow information to higher layers. The bidirectional feature pyramid network (BiFPN) [56] introduces a weighted feature fusion mechanism, assigning learnable weights to input features of different resolutions to mitigate fusion bias caused by scale discrepancies. The asymptotic feature pyramid network (AFPN) [57] enables direct interaction between non-adjacent layers and employs adaptive spatial fusion technology to bridge the cross-scale semantic gap. AL-YOLOv8, proposed by Zhang et al. [58], embeds adaptive spatial feature fusion and a Large-Kernel decoupled attention mechanism into the YOLOv8s detector head, achieving mAP@0.5 scores of 89.8%, 96.9%, and 92.2% on the DIOR, remote sensing object detection (RSOD), and NWPU VHR-10 datasets, respectively, while exhibiting robust false-positive suppression capabilities in complex backgrounds. Targeting the loss of

details during upsampling, CARAFE++ [59] introduces a content-aware feature reassembly mechanism that adaptively generates upsampled features with a controllable computational overhead. Residual asymptotic feature pyramid network (R-AFPN), proposed by Chen et al. [60], integrates three modules—residual asymptotic feature fusion, shallow information extraction, and hierarchical feature fusion—achieving 50.7% AP50 and 48.9% mAP50 on the TinyPerson and VisDrone2019 datasets, respectively, while reducing the parameter count by 15.1% compared to the baseline model.

Attention mechanisms have become one of the prominent research directions in small-object detection due to their unique advantages in feature selection and background suppression. The expanded attention module design of YOLO-PICO [61] achieves joint enhancement of spatial-channel features at a minimal computational cost, and its mixture-of-experts framework further boosts small-object-recognition performance. AAPW-YOLO, proposed by Wu et al. [62], adopts AKConv to substitute standard convolutions in the YOLOv8 backbone, thereby enhancing small-object-feature-extraction capabilities and combines a Wise-IoU gradient enhancement strategy to balance the gradient contributions of anchor boxes with varying qualities. Meng et al. [63] proposed a convolutional feature enhancement and attention fusion bidirectional feature pyramid network (BiFPN) framework, which enhances the global modeling capability of cross-scale feature fusion via BiFormer attention in synthetic aperture radar (SAR) ship-detection tasks. The global receptive field characteristic of the transformer architecture provides a technological pathway for small-object-feature modeling that is distinct from CNNs. Hyper-scale fusion (HSF)-DETR [64] constructs a hybrid progressive fusion network as the backbone to capture multi-scale information while preserving fine-grained details and combining position-aware aligned context and spatial tuning modules to achieve cross-scale feature calibration. The HyperFusion-DEIM framework, proposed by Liu et al. [65], integrates a multi-path attention network and scale-aware feature enhancement encoder, achieving 64.5% AP on the Small and Isolated Objects in Multi-view aerial images dataset (SIMD) with an inference throughput of 296.33 FPS. It outperforms RT-DETR and DEIM by 4.8% and 4.6%, respectively, thereby demonstrating a favorable trade-off between real-time efficiency and accuracy.

2.3. Application of the YOLO Series in Remote Sensing Image Object Detection

The YOLO series has emerged as one of the most widely applied single-stage frameworks in remote sensing object detection, owing to its end-to-end inference efficiency and continuously itemized performance optimization. Customized enhancements surrounding this series typically focus on three core dimensions: small-object-feature representation enhancement, multi-scale fusion optimization, and background interference suppression.

In recent years, improvement methodologies built upon YOLOv8 have proliferated steadily within remote sensing object detection. Wang et al. [66] introduced a bi-level routing attention mechanism into the high-level feature processing of the YOLOv8 backbone to suppress irrelevant background information, combining it with an asymptotic feature pyramid network to optimize multi-scale feature fusion. This approach achieved a mAP@0.5:0.95 score of 65.4% on the RSOD dataset, marking a 3.3 percentage point increase compared to the baseline model. The OA-YOLOv8, proposed by Zhang et al. [67], embeds an omni-dimensional perception refinement network into the backbone layer, enhancing multi-scale feature representation through the collaboration of receptive field convolution and a triplet attention mechanism; furthermore, it integrates DySample, the Haar wavelet transform, and a self-supervised equivalent attention mechanism during the upsampling stage to preserve fine-grained features, boosting mAP@0.5, precision, and recall on the SIMD dataset by 4.6%, 6.7%, and 4.1%, respectively. EM-YOLO [68] introduces a dynamic snake convolution to enhance feature extraction capabilities for rotating

objects and constructs a focusing-diffusion module within the feature fusion neck layer to effectively integrate multi-scale feature information. Supplemented by a layer-adaptive sparse pruning strategy to reduce the computational load, its deployment feasibility under resource-constrained scenarios was successfully validated on the Jetson Orin platform. The SPCI block design in YOLO-SPCI [69] incorporates selective flow gating, a perspective fusion module, and a class discrimination module. By embedding two SPCI blocks into the P3 and P5 stages of the YOLOv8 backbone, it significantly enhances the adaptive representation capability of multi-scale features and improves discriminative capability between classes.

In addition, substantial progress has been achieved in relevant research concerning other models within the YOLO series. Chen et al. [70] proposed a remote sensing image-detection method based on an improved YOLOv9 framework, which achieved superior performance by optimizing the feature extraction and bounding-box regression processes. Reference [71] introduced MSPA-YOLO to boost remote sensing object-recognition performance; by enhancing the architecture of YOLOv10, this method demonstrates strong competitiveness in terms of both detection accuracy and efficiency. The DroneScan-YOLO proposed by Bellec [72] tackles the triple challenges of small-object detection in drone imagery through a four-fold collaborative design scheme: enhancing input resolution, introducing a dynamic filter pruning mechanism, constructing a lightweight P2 detection branch, and designing a hybrid loss function. This method achieved a mAP@50 score of 55.3% on the VisDrone2019-DET dataset, marking a 16.6 percentage point increase over the YOLOv8s baseline mAP@50 and demonstrating improvements in recall rate; concurrently, it only increased the parameter count by 4.1% while maintaining an inference speed of 96.7 FPS.

Consequently, cross-architecture technology integration has become an important research direction in this field. DAE-YOLO [73] integrates a dynamic spatial sequence module into the YOLO11 architecture to enhance long-range spatial dependency modeling capabilities, an adaptive multi-scale feature enhancement module to achieve dynamic regulation of the receptive field, and an efficient bi-level attention mechanism, thereby achieving a certain advantage in real-time detection performance. YOLO-TLA [74] integrates a small-object-detection layer within the neck of YOLOv5 to boost the utilization efficiency of shallow-layer details. YOLO-DRS [75] introduces a lightweight multi-scale module (LEC) to enhance multi-scale feature fusion capabilities while controlling parameter growth. YOLO-MS [76] strengthens multi-scale feature representation through a multi-branch convolutional structure, which also achieves a certain competitive edge in terms of real-time detection performance.

The aforementioned studies have promoted the further application and development of the YOLO series in the field of remote sensing object detection; however, several key issues still remain unresolved: Spatial information loss induced by downsampling in deep networks has not been sufficiently mitigated; the degree of adaptivity in fusing features across different hierarchical levels remains limited. Feature separability between small objects and complex backgrounds requires further enhancement, and the trade-off between model lightweighting and detection accuracy is particularly significant in edge-deployment scenarios.

3. FFR-YOLO

Remote sensing scenarios are characterized by expansive coverage and complex backgrounds, where object discrimination often relies on abrupt edge transitions, corner responses, elongated contours, and subtle local differences rather than intense texture features. Although YOLOv8 offers robust real-time performance, its conventional convolutional downsampling tends to induce high-frequency aliasing and small-object edge loss,

while its neck's cross-scale fusion, designed for generic scenes, frequently leads to missed small objects, poor separation of dense objects, and increased false positives in complex backgrounds. To address these intertwined limitations, this study introduces a synergistic three-level optimization strategy within the YOLOv8 framework, covering the backbone, neck, and detector head front end. The resulting FFR-YOLO network comprises three core modules: FAPD, BRFM, and RDLGF, whose overall architecture is illustrated in Figure 1.

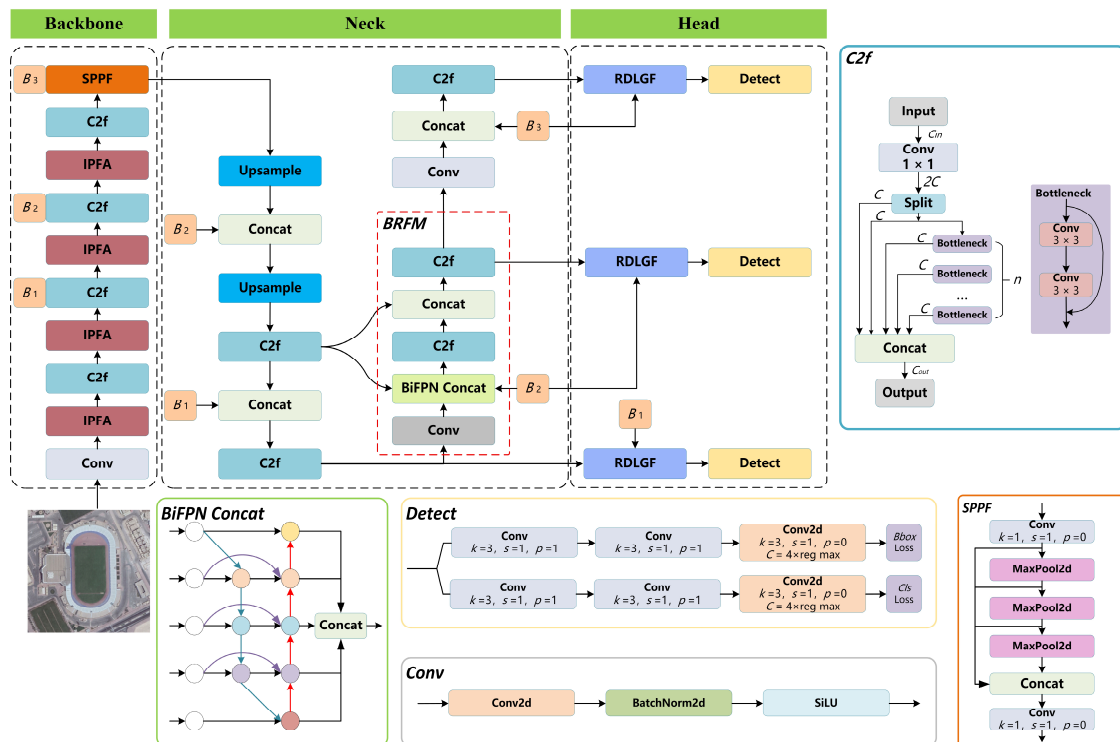


Figure 1. Overall network architecture of FFR-YOLO.

In the backbone, the FAPD module replaces strided convolutions with Haar wavelet decomposition to explicitly preserve low-frequency structures and directional high-frequency details during resolution compression. A low-frequency-guided gating mechanism selectively modulates high-frequency responses, mitigating aliasing and detail degradation while enhancing shallow-layer edge and contour preservation for subsequent propagation.

In the neck, the BRFM module reconstructs multi-scale fusion pathways through multi-source weighted aggregation and cross-pathway bridging interactions, enabling original backbone, top-down, and bottom-up features to undergo joint modeling at the same scale. This improves coordination between semantic abstraction and detail compensation, thereby stabilizing intermediate-layer representations for small and dense objects in complex backgrounds.

At the detector head front end, the RDLGF module performs recalibration and adaptive fusion of same-scale dual-path features from the backbone and neck. A joint global recalibration and spatial-channel gating mechanism dynamically estimates the relative contributions of local structural and global semantic information, enhancing classification and localization stability for elongated, adjacent, and background-embedded objects.

Functionally, these three architectural enhancements establish a cohesive, sequential relationship that seamlessly bridges preceding and succeeding computing stages. Operating at the front end of the backbone, FAPD preserves richer multi-frequency primitives, thereby laying a structurally clearer and information-rich foundation for the subsequent neck. The reconstructed neck channels these high-quality shallow- and middle-layer repre-

sentations across multiple interactive pathways, which significantly bolsters the stability of cross-scale feature propagation while expanding the representation capacity of the medium-scale semantic hub. RDLGF specifically redistributes the backbone features and fused features at the detector head input, enabling the structural advantages built up in the preceding two components to be effectively converted into enhanced classification and localization capabilities.

3.1. Bridge-Guided Bidirectional Reconstruction Fusion Module (BRFM)

The neck of YOLOv8 follows the FPN-PAN paradigm, propagating high-level semantics top-down and localization cues bottom-up. While effective for generic detection, this design exhibits two limitations in remote sensing scenarios. First, during bottom-up fusion, the network merely concatenates the current downsampled features with the corresponding top-down pathway features, neglecting the original same-scale backbone features and thus failing to establish multi-source collaboration. Second, although the two pathways intersect at specific nodes, they propagate independently thereafter, lacking continuous interaction and joint reconstruction at identical scales. This deficiency renders intermediate-scale features susceptible to semantic compression and texture interference, destabilizing the discriminative representations of small and dense objects.

To address these issues, this study constructs the BRFM within the neck to enhance the complementary utilization capability among features derived from heterogeneous sources, while improving information consistency during the multi-scale propagation process. Unlike existing FPN-, PAN-, and BiFPN-like structures, BRFM introduces original same-scale backbone features into each bottom-up fusion node and further incorporates bridge-guided reconstruction units between the top-down and bottom-up paths. This design enables same-scale features to undergo secondary interaction and recombination prior to the next propagation stage, thereby enhancing cross-path and cross-source collaborative reconstruction. The location of the BRFM within the overall network architecture is indicated by the red dashed box in Figure 1.

(1) Multi-source weighted fusion-based on BiFPN_Concat.

Conventional PAN nodes perform equal-weight concatenation on only two input paths, assuming equal contributions from different feature sources. This static approach struggles to accommodate the large-scale variations, complex backgrounds, and high object density typical of remote sensing imagery. To overcome this limitation, BiFPN_Concat weighted concatenation is introduced at each bottom-up fusion node, simultaneously aggregating three same-scale feature sources: the current bottom-up input, the corresponding top-down output, and the original backbone feature. Learnable weights are assigned to these heterogeneous representations, enabling dynamic adaptive fusion.

Let the three input features be denoted as F_{bu} , F_{td} , and F_{bk} ; here, F_{bu} represents the current input feature from the bottom-up pathway. F_{td} denotes the corresponding scale output feature from the top-down pathway, and F_{bk} represents the original same-scale feature provided by the backbone network.

First, learnable non-negative weights $\hat{w}_1, \hat{w}_2, \hat{w}_3$ are allocated to the three input paths, with all three weights initialized to 1 at the start of training, and a normalization process is subsequently applied to derive the fusion coefficients for each branch. Consequently, the weighted concatenation process of BiFPN_Concat can be formulated as Equations (1) and (2):

$$\hat{w}_i = \frac{w_i}{\sum_{j=1}^3 w_j + \epsilon}, i \in \{1, 2, 3\} \quad (1)$$

$$F_{mc} = \text{Concat}(\hat{w}_1 F_{bu}, \hat{w}_2 F_{td}, \hat{w}_3 F_{bk}) \quad (2)$$

where ε represents a small constant to prevent division by zero, $\text{Concat}(\cdot)$ denotes the concatenation operation along the channel dimension, and F_{mc} represents the intermediate feature obtained after the weighted multi-source concatenation. Subsequently, channel compression and local interactions are performed on the concatenation result through a convolutional mapping to obtain the fused output of the node as expressed by Equation (3):

$$F_{fuse} = \psi(F_{mc}) \quad (3)$$

where $\psi(\cdot)$ represents the feature transformation operation comprising convolution, normalization, and an activation function. Compared with direct equal-weight concatenation, BiFPN_Concat introduces learnable weights, thereby enabling the network to adaptively adjust the fusion intensity based on the effectiveness of features derived from different sources in various scenarios. In scenes with a relatively high proportion of small objects, detailed information from high-resolution pathways typically becomes more critical, allowing for the model to increase the weights of F_{td} or F_{bk} . Conversely, when the object scale is relatively large and relies on stronger semantic generalization, the network can appropriately enhance the contribution of low-resolution semantic features. Compared with the static combination methods under fixed topologies, this fusion strategy can, to a certain extent, adapt to the varying feature requirements arising from the content variations in remote sensing scenes.

(2) Bridge-guided bidirectional feature reconstruction.

While multi-source weighted aggregation mitigates information conflicts among heterogeneous features, it does not fully resolve the insufficient interaction depth between the top-down and bottom-up pathways. In the original neck, features from the two pathways converge only once at limited nodes and subsequently propagate independently, lacking continuous collaborative reconstruction between semantic guidance and detail compensation. To address this, bridge-guided fusion units are introduced between the two pathways to perform secondary joint modeling on same-scale intermediate features from both routes.

Let the output of the top-down pathway at scale s be denoted as $F_{td}^{(s)}$, and the intermediate feature of the bottom-up pathway at the corresponding scale as $F_{bu}^{(s)}$. BRFM first concatenates the two along the channel dimension to form the bridge input feature as expressed by Equation (4):

$$F_{br}^{(s)} = \text{Concat}(F_{td}^{(s)}, F_{bu}^{(s)}) \quad (4)$$

Subsequently, $F_{br}^{(s)}$ is fed into a C2f module for interactive reconstruction, yielding the bridge composite feature as expressed by Equation (5):

$$F_{out}^{(s)} = \text{C2f}(F_{br}^{(s)}) \quad (5)$$

where $F_{out}^{(s)}$ simultaneously fuses the high-level semantic information carried by the top-down pathway and the local structural details supplemented by the bottom-up pathway. Distinct from the direct output obtained via simple concatenation, the C2f module leverages branch aggregation and local residual connections to intensify the progressive interaction of cross-path information. This enables same-scale features to undergo more effective reconstruction and filtering before entering the subsequent propagation phase.

3.2. Frequency-Guided Anti-Alias Progressive Downsampling Module (FAPD)

Conventional strided convolutions in the YOLOv8 backbone, while efficient for natural scenes, tend to induce edge blurring and contour collapse in high-resolution remote sensing imagery where most objects occupy very few pixels, weakening the representation of small

and dense objects. To address this, the proposed FAPD replaces these convolutional downsampling layers, enabling controlled information reorganization and selection during resolution compression. In contrast to existing wavelet-based downsampling or anti-aliasing sampling methods, FAPD introduces a low-frequency-guided high-frequency gating mechanism that performs unified channel reorganization and linear mapping after subband decomposition. By leveraging the LL subband to selectively filter the effectiveness of LH, HL, and HH components, FAPD establishes an adaptive balance between detail fidelity and background noise suppression. Figure 2 illustrates the architectural diagram of the FAPD module.

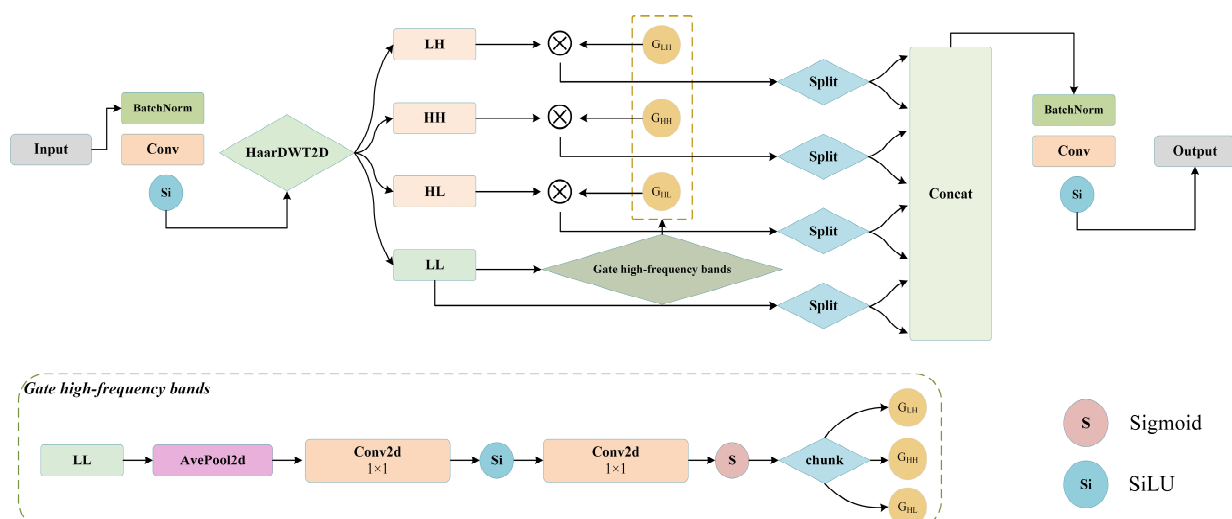


Figure 2. An architectural diagram of the FAPD module.

During the process of feature compression on remote sensing imagery, convolutional downsampling layers are highly susceptible to structural degradation. To address these issues, the proposed module is deployed at multiple downsampling nodes within the backbone, thereby preserving richer structural information throughout the feature compression process across diverse scales. The module hierarchically comprises five essential components: pre-convolution, wavelet decomposition, high-frequency gating, channel reorganization, and linear fusion. Upon entering the module, the input feature $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$ first passes through a preprocessing branch comprising a 3×3 convolution, normalization, and a SiLU activation function, thereby yielding the locally aggregated feature F_{pre} . This process can be expressed as Equation (6):

$$F_{pre} = \phi(BN(Conv_{3 \times 3}(X))) \quad (6)$$

where ϕ denotes the SiLU activation function. This step performs a learnable blending operation within the local neighborhood before the downsampling is executed. Following the pre-aggregation, the information within the neighborhood is initially integrated, which stabilizes the subsequent low-frequency and high-frequency responses obtained via frequency decomposition, thereby alleviating the aliasing artifacts that might be induced by direct sampling.

Figure 3 shows a schematic diagram of the HaarDWT2D structure. The core of this module lies in replacing traditional strided convolutions with Haar wavelet decomposition to accomplish downsampling. Specifically, for each 2×2 neighborhood, the pixels x_{00} , x_{01} , x_{10} , x_{11} are obtained via pixel extraction. Subsequently, four subbands are constructed, where the expressions for the low-frequency approximation component low-low (LL) and

the three high-frequency detail components (low-high (*LH*), high-low (*HL*), and high-high (*HH*)) are formulated as Equation (7):

$$\begin{aligned} LL &= \frac{1}{2}(x_{00} + x_{01} + x_{10} + x_{11}), \\ LH &= \frac{1}{2}(x_{00} - x_{01} + x_{10} - x_{11}), \\ HL &= \frac{1}{2}(x_{00} + x_{01} - x_{10} - x_{11}), \\ HH &= \frac{1}{2}(x_{00} - x_{01} - x_{10} + x_{11}). \end{aligned} \quad (7)$$

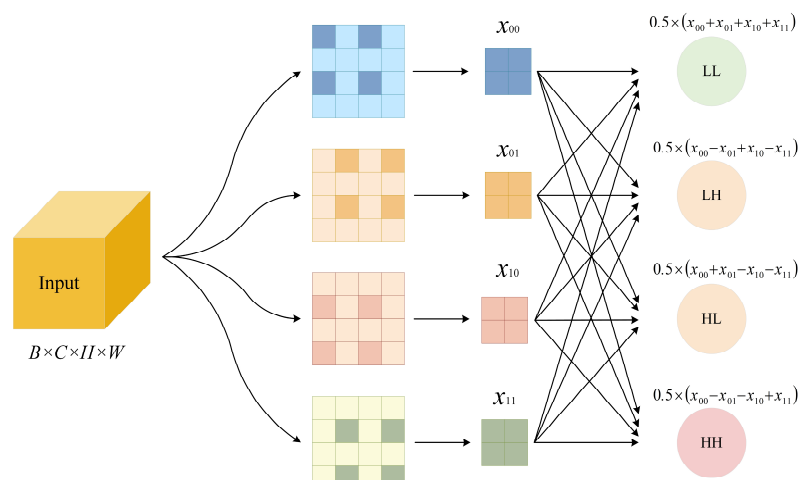


Figure 3. A structural diagram of HaarDWT2D.

LL carries the low-frequency contour and coarse semantic information of the scene, while *LH*, *HL*, and *HH* capture directional high-frequency details. Unlike strided convolution that jointly performs sparse sampling and weighted aggregation, wavelet decomposition explicitly separates each 2×2 neighborhood into orthogonal subbands, preserving structural details as explicit tensors for deeper layers.

However, retaining all high-frequency components indiscriminately is insufficient, as background regions in remote sensing imagery generate meaningless regular textures that appear as large-amplitude repetitive patterns, potentially misleading the detector. To address this, a low-frequency-guided high-frequency gating mechanism is introduced after wavelet decomposition. Using the *LL* subband, which condenses global scene distribution, as contextual reference, the gating branch applies global average pooling and two 1×1 convolutions to generate channel-wise gating coefficients g_{LH} , g_{HL} , g_{HH} , as expressed by Equation (8):

$$[g_{LH}, g_{HL}, g_{HH}] = \sigma(\text{GAP}(LL)) \quad (8)$$

where σ represents the Sigmoid function. The gating coefficients g_{LH} , g_{HL} , and g_{HH} are split along the channel dimension into three independent gating vectors, which are then multiplied with the three high-frequency subbands in a channel-wise manner, respectively, as expressed by Equation (9):

$$\begin{aligned} \widehat{LH} &= LH \odot g_{LH}, \\ \widehat{HL} &= HL \odot g_{HL}, \\ \widehat{HH} &= HH \odot g_{HH}, \end{aligned} \quad (9)$$

$\odot$ denotes the channel-wise multiplication, and the value of the gating coefficient ranges between 0 and 1. If the low-frequency pooled features capture clear object structures, the corresponding channel gate approaches 1 and preserves high-frequency details. In contrast,

for homogeneous backgrounds or strong random noise, the gate is suppressed to prevent unreliable high-frequency responses from entering feature fusion. Thus, high-frequency retention is determined not merely by response amplitude, but by the contextual constraints provided by low-frequency structural information. Since low-frequency components encode the global scene layout, they provide a reliable basis for assessing high-frequency credibility: structure-consistent edge responses are retained, whereas responses from background textures or repetitive patterns are attenuated. This enables scene-adaptive frequency filtering, reducing invalid high-frequency interference while enhancing salient object boundaries.

Following the frequency filtering, all four subbands, $LL, \widehat{LH}, \widehat{HL}, \widehat{HH}$, are halved and rearranged along the channel dimension. Subsequently, the eight feature segments are interleaved and concatenated into a 4C-channel tensor, enabling the proposed module to directly replace the convolutional downsampling layers in YOLOv8 without altering subsequent network interfaces. Rather than being forwarded as independent low-frequency, directional high-frequency, and diagonal components, these subbands are reorganized into finer-grained channel segments, allowing the model to jointly capture object contours, local edges, and background-suppression cues. Finally, a 1×1 convolution performs cross-frequency fusion and channel compression, mapping the reorganized 4C features to the target dimension. Compared with conventional convolutional downsampling, the resulting features preserve backbone low-frequency information more effectively while reducing the influence of meaningless high-frequency noise on deeper representations.

3.3. Recalibrated Dual-Branch Local–Global Fusion (RDLGF)

Traditional feature pyramid pathways usually fuse multi-scale features by element-wise addition or channel-wise concatenation followed by standard convolutions, which implicitly assumes equal and static contributions from different sources. In remote sensing scenes, however, large-scale variations and complex backgrounds often cause such fixed-weight fusion to suppress local textures in noisy regions or introduce redundant global interference in sparse object areas. To address this issue, RDLGF is proposed to dynamically adjust the contribution of each branch according to the input feature content, thereby adaptively enhancing the fused representation in both spatial and channel dimensions. In contrast to prevalent single-branch attention-based recalibration approaches such as SE and CBAM, RDLGF distinguishes itself by first conducting joint recalibration on the same-scale features derived from both the backbone and neck branches. Subsequently, it generates spatial-channel decoupled gating guided by the shared feature representation and finally accomplishes dynamic allocation of the two feature streams via a complementary fusion strategy.

Figure 4 shows a schematic diagram of the RDLGF module structure. From the implementation perspective, RDLGF employs three convolutional structures. ConvBNAct, consisting of convolution, batch normalization, and SiLU activation, is used for spatial feature learning at the original resolution $H \times W$, where batch normalization stabilizes feature distributions and improves convergence. ConvAct removes batch normalization and is mainly applied to the 1×1 channel descriptors generated after global average pooling, ensuring stable channel-wise responses. DWConvBNAct adopts depth-wise separable convolution for lightweight feature refinement, reducing parameters and computation while maintaining representation capability.

The RDLGF module is deployed at the front end of detector heads to replace the simplified feature concatenation operations in the original architecture. Its input receives isometric feature tensors from two parallel forward propagation pathways: one, $x_b, x_n \in \mathbb{R}^{B \times C \times H \times W}$, x_b is same-scale direct-connection features from the backbone; x_n , the other, is same-scale features reconstructed and fused from the neck.

The overall operational pipeline of the module can be partitioned into three sequential stages: the joint recalibration and guidance generation, spatial-channel decoupled gating, and complementary fusion. Structurally, RDLGF can be divided into three parts: the joint recalibration after concatenation, fusion-guided feature generation, and local-global collaborative gating. The inputs to the module consist of two feature maps, denoted as $x_b, x_n \in \mathbb{R}^{B \times C \times H \times W}$, which possess completely identical geometric shapes. The module initially concatenates these inputs along the channel dimension to yield $z = [x_b, x_n] \in \mathbb{R}^{B \times 2C \times H \times W}$. This step effectively preserves the complete response distributions of both input pathways, thereby enabling the subsequent weight generation process to comprehensively evaluate the relative contributions of the two features within a joint representation space.

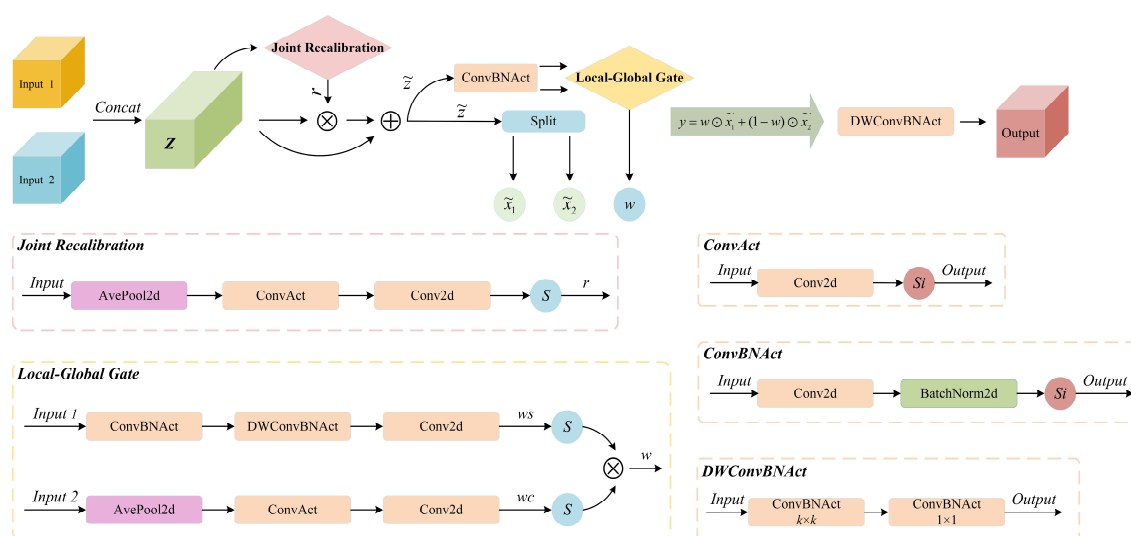


Figure 4. A structural diagram of the RDLGF module.

Therefore, the first stage of RDLGF performs joint recalibration to reduce channel-level distribution shifts between the two streams and generate a guidance coefficient r for subsequent gating. Specifically, the concatenated feature z is compressed by global average pooling into a channel descriptor. A lightweight mapping composed of two 1×1 convolutions is then used to produce a $2C$ -dimensional weight vector, which is normalized by a Sigmoid function to obtain the channel-wise recalibration coefficient r , as formulated in Equation (10).

$$r = \sigma(W_2 \delta(W_1 \text{GAP}(z))), \quad (10)$$

Herein, GAP denotes the global average pooling operation; $\delta(\cdot)$ represents the SiLU activation function, and $\sigma(\cdot)$ represents the Sigmoid function. In addition, w_1 and w_2 correspond to the weighted parameters of the two consecutive point-wise convolutions, respectively. Upon obtaining the modulation vector, the aggregated features are recalibrated in a residual manner, which can be expressed as Equation (11):

$$\tilde{z} = z \odot r + z. \quad (11)$$

where $\odot$ denotes the element-wise multiplication. The objective of this design is to enable the network to adaptively suppress interference from less informative channels based on global statistical properties while maintaining the stability of gradient propagation through the identity mapping branch. Rather than directly learning the fusion weights from the original x_b, x_n , this branch first allows for the network to distill a more stable criterion

for judgment within the blended representation after joint recalibration. Consequently, the fusion weights no longer depend on the local deviations of any single input pathway but are instead established upon the outcome of the collaborative analysis between the dual branches. Following the joint recalibration, the module splits $\tilde{z}$ along the channel dimension back into $\tilde{x}_b, \tilde{x}_n$. Concurrently, a 1×1 ConvBNAct convolutional layer is employed to compress the $2C$ channels back to C channels, thereby yielding the guidance feature M that serves as the input source for the dual-gating branches.

The key design of RDLGF lies in the spatial-channel decoupled gating stage. Rather than being constrained to a single modality of conventional channel or spatial attention, this stage utilizes the guidance feature M to generate a weight field that shares the identical dimensions as the features to be fused, possessing both spatial continuity and channel specificity. The gating network is split into two parallel substructures for spatial-dimension and channel-dimension gating. Spatial gating is tasked with localizing the regions that are more likely to correspond to objects, whereas channel gating emphasizes the more credible semantic responses. In remote sensing imagery recognition, relying solely on spatial attention may cause the model to allocate excessive focus to high-response background areas; conversely, relying exclusively on channel attention makes it difficult to distinguish the true validity of the same semantic channel across different locations. By implementing a multiplicative collaborative method between the spatial and channel dimensions, a position-semantic joint constraint is established, thereby significantly enhancing the feature selection capability of the network within complex scenarios. The spatial gating branch takes the fusion guidance feature as input, initially utilizing a 1×1 convolution to accomplish mapping and subsequently extracting local neighborhood context via a depth-wise convolution to ultimately generate the spatial response map $W_s \in \mathbb{R}^{B \times 1 \times H \times W}$. Concurrently, the channel gating branch compresses the spatial dimensions leveraging global average pooling to estimate the significance of each class of response along the channel dimension, ultimately generating the channel response map $W_c \in \mathbb{R}^{B \times C \times 1 \times 1}$. The composite weight tensor $W \in \mathbb{R}^{B \times C \times H \times W}$, which is applied to the feature fusion at the last stage, is synthesized through the broadcast multiplication of the Sigmoid-normalized outputs from both branches as expressed by Equation (12):

$$w = \sigma(w_s) \odot \sigma(w_c), \quad (12)$$

Building upon this, the module adopts a complementary fusion strategy rather than simple addition or concatenation-followed-by-convolution. The final output y is formulated as Equation (13):

$$y = w \odot \tilde{x}_b + (1 - w) \odot \tilde{x}_n \quad (13)$$

When the gating network assigns a larger weight to a specific channel at a certain spatial location, the fused output relies more on the local-detail branch $\tilde{x}_b$; when the weight decreases, the output shifts toward the semantic-context branch $\tilde{x}_n$. In this way, RDLGF enables adaptive feature allocation across spatial regions and semantic channels. The backbone branch mainly preserves original hierarchical structures and fine local details, while the neck-reconstructed branch provides semantic supplementation and scale-aware context. This adaptive selection is particularly useful for remote sensing detection. For small and densely distributed objects under complex backgrounds, the fusion weights tend to retain branches with clearer textures, contours, edges, and local contrast. For larger targets or objects requiring contextual discrimination, the weights shift toward branches with more stable semantic representations. Thus, the module can flexibly balance local detail preservation and global semantic enhancement according to object scale and scene complexity. Although gating fusion performs effective information selection, it may intro-

duce local discontinuities in spatial consistency and neighborhood smoothness. Therefore, a depth-wise convolution is further applied at the end of the module to recover local contextual correlations, followed by a point-wise convolution for cross-channel reintegration. This refinement improves the structural integrity of the fused features with only a small computational cost.

4. Experiment

4.1. Experimental Datasets

To comprehensively validate the effectiveness of the proposed FFR-YOLO in optical remote sensing object detection tasks, two publicly available and widely adopted benchmark datasets were selected for this study, namely, the object detection in optical remote sensing images (DIOR) dataset [77] and the Northwestern polytechnical university very high resolution 10-class (NWPU VHR-10) dataset [78]. These benchmarks distinctively vary in terms of image quantities, spatial resolutions, target classes and size distributions, scenario diversities, and annotation granularities, thereby offering a multi-dimensional and systematic evaluation of the generalization capability and robustness of the detection models. In this study, instead of using the official predefined splits, we adopted a random 7:3 train-validation split ratio for both datasets, with a category-balanced stratification strategy that ensured each object class maintained a 7:3 distribution between the training set and the validation set. Figure 5 illustrates the instance counts per class and the object size distribution statistics for both datasets. It should be noted that as this random split differs from the official predefined partitions, the absolute mAP values reported in our experiments may not be directly numerically comparable with those published in studies adopting the official splits. Nonetheless, because all competing models were trained and evaluated under this identical partition, the relative improvements demonstrated by FFR-YOLO over its baselines remain valid and statistically fair.

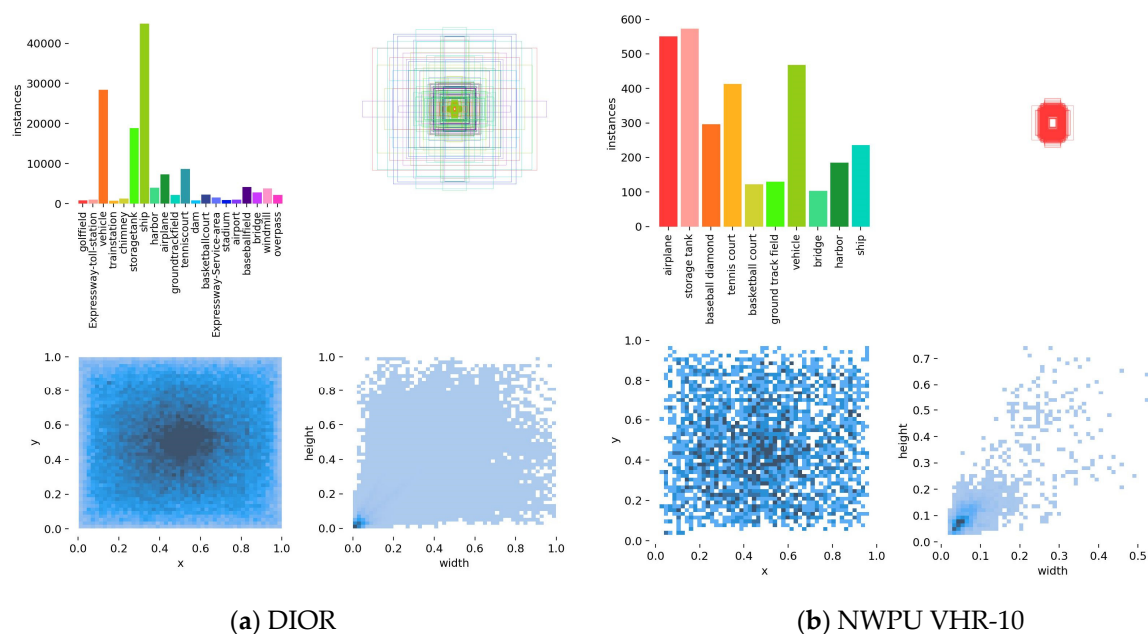


Figure 5. Data distribution: (a) DIOR dataset; (b) NWPU VHR-10 dataset.

The DIOR dataset, proposed by Northwestern Polytechnical University, is a large-scale benchmark for object detection in optical remote sensing images. It contains 23,463 images and 192,472 annotated instances, covering 20 object classes: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll

gate, harbor, golf course, ground track field, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and wind mill. The images in this dataset were collected from more than 80 regions across the globe, spanning diverse seasons, weather conditions, and imaging parameters. The spatial resolution ranges from 0.5 to 30 m, resulting in significant challenges associated with high intra-class variations and high inter-class similarities.

The NWPU VHR-10 dataset, also constructed by Northwestern Polytechnical University, is a medium-scale, high-resolution optical remote sensing benchmark for object detection. It consists of 800 optical remote sensing images with a spatial resolution spanning from 0.5 to 2 m. Among these, 650 images contain targets, while 150 images contain no targets. The dataset annotates a total of 3651 instances, encompassing 10 geospatial object classes: airplane, ship, storage tank, baseball field, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. All targets are manually annotated using horizontal bounding boxes. Although this dataset is relatively small in scale, its high image resolution and diverse target scales make it widely utilized for the rapid validation of remote sensing object-detection algorithms and the evaluation of small-target detection performance.

4.2. Experimental Setup

The experiments in this study were implemented based on the PyTorch deep learning framework, utilizing YOLOv8 as the baseline network architecture for subsequent modifications and training, with the Ultralytics implementation version 8.2.0. The experiments ran on a 64-bit Windows 10 operating system. Hardware configuration included an Intel Core i9 series CPU, paired with a single NVIDIA GeForce RTX 3090 GPU equipped with 24 GB of VRAM to accelerate both the model training and inference processes. Python 3.8 was selected as the programming language, with PyTorch 2.0.0 and CUDA 11.6 establishing the foundational framework for deep learning and GPU acceleration. PyCharm 2022 was employed as the integrated development environment. Furthermore, to eliminate the interference of pre-trained weights and to precisely quantify the performance gains attributable to each proposed module, all experiments were conducted without any pretrained weights.

The hyperparameter settings for model training are detailed in Table 1.

Table 1. Hyperparameter configurations for model training.

Parameters	Setup
Epochs	150
Batch size	12
Workers	8
Initial learning rate	0.001
Optimizer	AdamW
Random seed	96
Input image size	800 × 800
mixup	0.4
mosaic	1.0

4.3. Evaluation Metrics

To objectively and comprehensively evaluate the performance of the proposed FFR-YOLO network in remote sensing object-detection tasks, a series of evaluation metrics were selected from two aspects: detection accuracy and model complexity.

In terms of detection accuracy, precision (P), recall (R), average precision (AP), and mean average precision (mAP), which are universally adopted in the field of object detection, were employed as the core evaluation metrics. Let true positive (TP) denote the number of correctly detected positive samples, false positive (FP) the number of negative samples incorrectly

detected as targets, and false negative (FN) the number of undetected positive samples. The mathematical formulas for precision and recall are defined by Equations (14) and (15):

$$\text{Precision} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

AP measures the area under the precision–recall curve for a single class, whereas the mAP is computed as the arithmetic mean of the AP values across all classes. Furthermore, this study concurrently employed two specific metrics, namely, mAP@0.5 and mAP@0.5:0.95, to reflect the detection capability of the model under loose and stringent localization constraints, respectively, as expressed by Equation (16):

$$AP = \int_0^1 P(R) dR \quad (16)$$

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (17)$$

In terms of model complexity, parameter count, params, and giga floating-point operations (GFLOPs) were adopted as the quantitative metrics. Specifically, parameter count was utilized to evaluate the storage space requirements of the model, while GFLOPs represented the billion-level floating-point operations required for a single forward inference. Collectively, these two metrics reflected the overall resource consumption levels of the model during deployment. To further assess the practical inference efficiency, frames per second (FPS) was additionally reported as a complementary metric for real-time applicability.

To ensure the statistical reliability of our experimental results, all reported metrics were obtained by averaging over five independent runs. For each run, the model was trained from scratch with the same hyperparameter configuration. It is worth noting that the standard deviations of all reported metrics across the five runs were consistently below 0.4%, confirming the stability of our training process and the statistical reliability of the reported improvements.

4.4. Ablation Experiments

4.4.1. Comprehensive Ablation Experiments

To verify the individual and collective effectiveness of the three proposed components (i.e., BRFM, FAPD, and RDLGF), ablation experiments were conducted on both the DIOR and NWPU VHR-10 datasets. Utilizing the baseline YOLOv8 as the anchor, different combinations of these modules were integrated into the network to investigate their mutual impacts and collaborative relationships. The quantitative ablation results evaluated on the DIOR and NWPU VHR-10 datasets are summarized in Tables 2 and 3, respectively, where the checkmark ✓ indicates the inclusion of the corresponding module, whereas the × signifies its exclusion.

Table 2. Ablation experiments based on the DIOR dataset.

YOLOv8	BRFM	FAPD	RDLGF	P (%)	R (%)	mAP ₅₀ (%)	mAP _{50:95} (%)	Param (M)	FLOPs	FPS
✓	×	×	×	86.5	77.7	83.6	60.1	3.0	8.1	163
✓	✓	×	×	86.1	78.5	84.0	60.5	3.1	8.6	158
✓	×	✓	×	87.1	78.9	84.8	61.6	3.7	11.4	151
✓	×	×	✓	86.8	78.1	84.1	60.9	3.4	8.8	147
✓	✓	✓	×	88.1	78.4	84.8	62.0	3.9	11.9	149
✓	✓	×	✓	87.5	78.3	84.6	61.8	3.5	9.2	175
✓	×	✓	✓	86.9	78.2	84.5	62.3	4.2	12.2	149
✓	✓	✓	✓	88.5	79.4	85.8	63.1	4.3	12.6	136

Table 3. Ablation experiments based on the NWPU VHR-10 dataset.

YOLOv8	BRFM	FAPD	RDLGF	P (%)	R (%)	mAP ₅₀ (%)	mAP _{50:95} (%)	Param (M)	FLOPs	FPS
✓	×	×	×	90.3	83.2	88.3	56.1	3.0	8.1	90
✓	✓	×	×	91.3	84.4	90.7	59.8	3.1	8.6	104
✓	×	✓	×	93.4	85.3	92.6	59.1	3.7	11.4	89
✓	×	×	✓	92.8	87.0	92.6	60.2	3.4	8.8	95
✓	✓	✓	×	93.3	87.2	92.8	60.1	3.9	11.9	89
✓	✓	×	✓	95.5	83.9	92.8	60.3	3.5	9.2	90
✓	×	✓	✓	93.2	83.4	92.3	60.7	4.2	12.2	86
✓	✓	✓	✓	93.6	88.9	93.6	62.1	4.3	12.6	81

The ablation results indicated that all three proposed modifications yielded consistent performance gains across both datasets, albeit with distinctly different strategic emphasis. On the DIOR dataset, after introducing BRFM alone, mAP₅₀ and mAP_{50:95} increased from 83.6% and 60.1% to 84.0% and 60.5%, respectively, with an additional cost of only 0.1 M parameters and 0.5 GFLOPs, indicating that neck-path reconstruction can alleviate information fragmentation during multi-scale feature transmission, although the improvement remains relatively moderate. When the FAPD was configured independently, the mAP₅₀ and mAP_{50:95} increased to 84.8% and 61.6%, respectively. This represents the most significant margin of growth among the individual modules, albeit at the highest individual overhead of 0.7 M parameters and 3.3 GFLOPs, which underscores the definitive advantages of frequency-domain anti-aliasing downsampling in terms of contour preservation for remote sensing small targets, detail retention, and complex background suppression. When RDLGF was applied independently, mAP₅₀ and mAP_{50:95} increased to 84.1% and 60.9%, respectively, with a moderate overhead of 0.4 M parameters and 0.7 GFLOPs, demonstrating that dual-branch recalibration before detection helped enhance classification and localization stability. Upon the joint integration of all three modules, the precision, recall, mAP₅₀, and mAP_{50:95} reached 88.5%, 79.4%, 85.8%, and 63.1%, respectively, which exhibited increments of 2.0, 1.7, 2.2, and 3.0 percentage points over the baseline. The complete model incurred a total overhead of 1.3 M parameters and 4.5 GFLOPs, with FPS decreasing from 163 to 136. This reinforced that the three components possessed a strong intrinsic complementarity across different stages.

On the NWPU VHR-10 dataset, the performance enhancements achieved by each individual module were even more significant. Specifically, the independent introduction of the BRFM increased the mAP₅₀ from 88.3% to 90.7% and boosted the mAP_{50:95} by 3.7 percentage points. Notably, BRFM achieved this with minimal overhead (0.1 M parameters, 0.5 GFLOPs) and even increased FPS from 90 to 104, attributed to its efficient feature reuse strategy. The FAPD further propelled the mAP₅₀ to 92.6%, which confirmed its specific advantage in preserving fine-grained structural details within high-resolution remote sensing imagery, though it added 0.7 M parameters and 3.3 GFLOPs, reducing FPS from 90 to 89. The application of RDLGF increased the precision to 92.8% and pushed the mAP_{50:95} to 60.2%, reflecting its positive role in false-alarm suppression and bounding-box regression alignment, with a moderate overhead of 0.4 M parameters and 0.7 GFLOPs. The complete model finally achieved a mAP₅₀ of 93.6% and a mAP_{50:95} of 62.1%, outperforming the baseline by 5.3 and 6.0 percentage points, respectively. Meanwhile, the total overhead reached 1.3 M parameters and 4.5 GFLOPs, with FPS decreasing from 90 to 81. The results demonstrated that the proposed method achieved favorable performance gains with limited extra overhead and maintained real-time inference capability across both datasets.

To evaluate the sensitivity of the FAPD module to different wavelet bases, we compared the default Haar wavelet against Daubechies-4 and Symlets-4. As presented in

Table 4, Haar consistently achieved the highest mAP50 on both DIOR (85.8%) and NWPU VHR-10 (93.6%), marginally surpassing db4 by 0.9% and 1.6% and sym4 by 1.2% and 2.0%, respectively. In terms of mAP50:95, Haar also maintained a leading position, outperforming db4 and sym4 by 1.3% and 1.6% on DIOR and 2.4% and 2.3% on NWPU VHR-10. Notably, although db4 and sym4 achieved slightly higher FPS on the NWPU VHR-10 dataset, this was primarily attributed to the smaller image resolution and reduced computational workload of that dataset; on the larger DIOR dataset, Haar remained the fastest (136 FPS) due to its shortest filter length and simplest computation. The performance gaps among the three wavelets were marginal (<2% in mAP50), indicating that FAPD was robust to wavelet selection. Given Haar's best overall accuracy-efficiency trade-off and its computational simplicity, we retained it as the default choice in FAPD.

Table 4. The sensitivity analysis of different wavelet types in FAPD on the DIOR and NWPU VHR-10 dataset.

	Wavelet Type	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	FPS
DIOR	Haar	88.5	79.4	85.8	63.1	136
	Daubechies-4	87.6	78.5	84.9	61.8	130
	Symlets-4	87.2	78.0	84.6	61.5	128
NWPU VHR-10	Haar	93.6	88.9	93.6	62.1	81
	Daubechies-4	95.4	84.4	92.0	59.7	87
	Symlets-4	93.9	84.8	91.6	59.8	90

4.4.2. Comparison of Downsampling Methods

To further evaluate the isolated effectiveness of the proposed FAPD, it was comprehensively compared against several state-of-the-art downsampling approaches, including three specialized modules: ADown [47], EUCB [79], and IPFA [80]. To ensure a rigorous control of variables, the different downsampling strategies were consistently deployed at the four structural downsampling nodes within the backbone network. The quantitative experimental data obtained on the two benchmark datasets are summarized in Table 5.

Table 5. Comparative experiments of different downsampling modules.

	Methods	P (%)	R (%)	mAP ₅₀ (%)	mAP _{50:95} (%)	Param (M)	FLOPs	FPS
DIOR	Baseline	86.5	77.7	83.6	60.1	3.0	8.1	163
	ADown	86.8	78.1	84.0	60.6	2.7	7.7	174
	EUCB	86.9	78.4	84.3	61.0	3.0	8.5	156
	IPFA	86.7	78.8	84.2	61.2	3.0	9.7	164
	FAPD	87.1	78.9	84.8	61.6	3.7	11.4	136
NWPU VHR-10	Baseline	90.3	83.2	88.3	56.1	3.0	8.1	90
	ADown	91.7	83.9	90.8	57.6	2.7	7.7	103
	EUCB	92.5	84.7	92.1	58.6	3.0	8.5	98
	IPFA	91.6	87.4	91.2	58.9	3.0	9.7	92
	FAPD	93.4	85.3	92.6	59.1	3.7	11.4	89

The results in Table 5 demonstrate that FAPD achieved optimal or near-optimal accuracy on both datasets. On DIOR, the baseline obtained 83.6% mAP50 and 60.1% mAP50:95. Although ADown, EUCB, and IPFA provided modest improvements, their overall performance remained below that of FAPD. ADown prioritized efficiency, achieving the highest speed of 174 FPS while reducing parameters and computational complexity, whereas EUCB and IPFA improved mAP50:95 to 61.0% and 61.2%, respectively. In contrast, FAPD achieved 84.8% mAP50 and 61.6% mAP50:95, with precision and recall reaching 87.1% and 78.9%. Its speed decreased to 136 FPS, 16.6% below the baseline, indicating that FAPD traded moderate inference efficiency for improved object discrimination and bounding-box regression.

The advantage of FAPD was more pronounced on NWPU VHR-10, where it achieved the highest precision (93.4%) and mAP50 (92.6%). Its mAP50:95 reached 59.1%, exceeding the baseline by 3.0 percentage points and outperforming the other downsampling methods. Although ADown remained the fastest at 103 FPS, FAPD operated at 89 FPS, nearly matching the baseline (90 FPS). Despite a moderate increase in parameters and FLOPs, this limited overhead was acceptable given its consistent gains across multiple accuracy metrics. These results indicated that frequency-domain decomposition and low-frequency-guided gating effectively enhance structural fidelity and anti-aliasing capability for remote sensing object detection.

4.4.3. Comparison of Feature Fusion Modules

To further evaluate the effectiveness of the proposed RDLGF module, a comprehensive comparison was conducted against three distinct feature fusion modules, namely, SDFM [81], LGAG [82], and BiFPN_Concat [22]. To minimize variable interference, different fusion strategies were separately placed at the front end of the detector head for feature fusion. The experimental results of each module on the two datasets are shown in Table 6.

Table 6. Comparative experiments of different feature fusion modules.

	Methods	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	Param (M)	FLOPs	FPS
DIOR	Baseline	86.5	77.7	83.6	60.1	3.0	8.1	163
	SDFM	86.6	77.9	83.9	60.6	3.4	8.8	169
	LGAG	86.9	77.8	83.6	60.5	5.3	12.1	176
	BiFPN_Concat	87.0	77.9	84.0	60.7	4.3	13.4	171
	RDLGF	86.8	78.1	84.1	60.9	3.4	8.8	136
NWPU VHR-10	Baseline	90.3	83.2	88.3	56.1	3.0	8.1	90
	SDFM	91.3	84.9	90.8	59.9	3.4	8.8	94
	LGAG	93.3	85.1	91.8	59.6	5.3	12.1	100
	BiFPN_Concat	94.1	85.7	92.4	59.7	4.3	13.4	97
	RDLGF	92.8	87.0	92.6	60.2	3.4	8.8	89

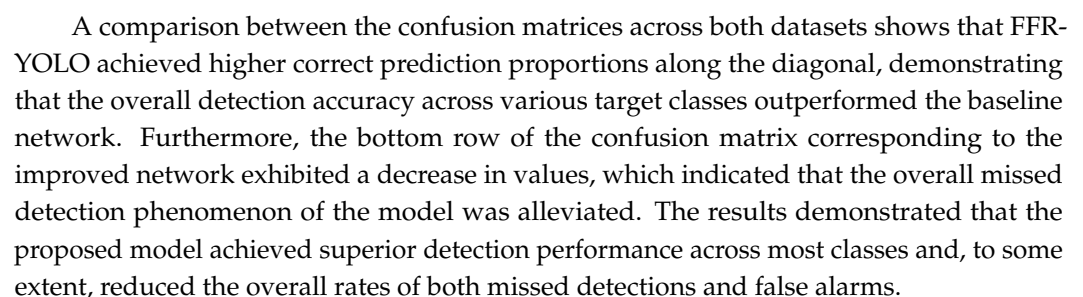
The results in Table 6 show that RDLGF achieved the best overall feature fusion performance on both datasets. On DIOR, SDFM improved mAP50 and mAP50:95 from 83.6% and 60.1% to 83.9% and 60.6%, respectively, whereas LGAG increased precision to 86.9% without improving mAP50. BiFPN_Concat achieved 84.0% mAP50 and 60.7% mAP50:95 but incurred higher computational costs. In comparison, RDLGF added only 0.4 M parameters and achieved the highest mAP50 (84.1%) and mAP50:95 (60.9%), with recall reaching 78.1%. Although its inference speed decreased to 136 FPS, these results confirmed the effectiveness of RDLGF in feature recalibration and detection accuracy enhancement.

On NWPU VHR-10, RDLGF obtained the highest mAP50 (92.6%), mAP50:95 (60.2%), and recall (87.0%), indicating improved detection of dense, small, and adjacent objects. Its inference speed of 89 FPS was also close to that of the baseline (90 FPS) despite its slightly increased computational cost. Although LGAG and BiFPN_Concat achieved higher precision, their greater complexity did not result in superior mAP50:95. Overall, RDLGF effectively integrated joint recalibration, spatial-channel decoupled gating, and complementary fusion, thereby balancing detection accuracy, recall, inference efficiency, and model complexity.

4.5. Comparative Analysis with YOLOv8

Figure 6 illustrates the confusion matrices before and after the proposed modifications, where (a) and (c) correspond to the confusion matrices of the original baseline model on the NWPU VHR-10 and DIOR datasets, respectively, while (b) and (d) represent those of

Figure 6. Confusion matrix comparison: (a) The confusion matrix of YOLOv8 on the NWPU VHR-10 dataset; (b) the confusion matrix of FFR-YOLO on the NWPU VHR-10 dataset; (c) the confusion matrix of YOLOv8 on the DIOR dataset; (d) the confusion matrix of FFR-YOLO on the DIOR dataset.



To evaluate the effectiveness and convergence characteristics of the improved FFR-YOLO network, we compared the box loss (box_loss), classification loss (cls_loss), and distribution focal loss (dfl_loss) of original YOLOv8 and the proposed FFR-YOLO during training, with results shown in Figure 7. Across the entire 150-epoch training cycle, FFR-YOLO's three core loss curves stayed lower in most training stages and converged to smaller values in the later phase, proving its advantages in convergence efficiency and feature representation learning during optimization.

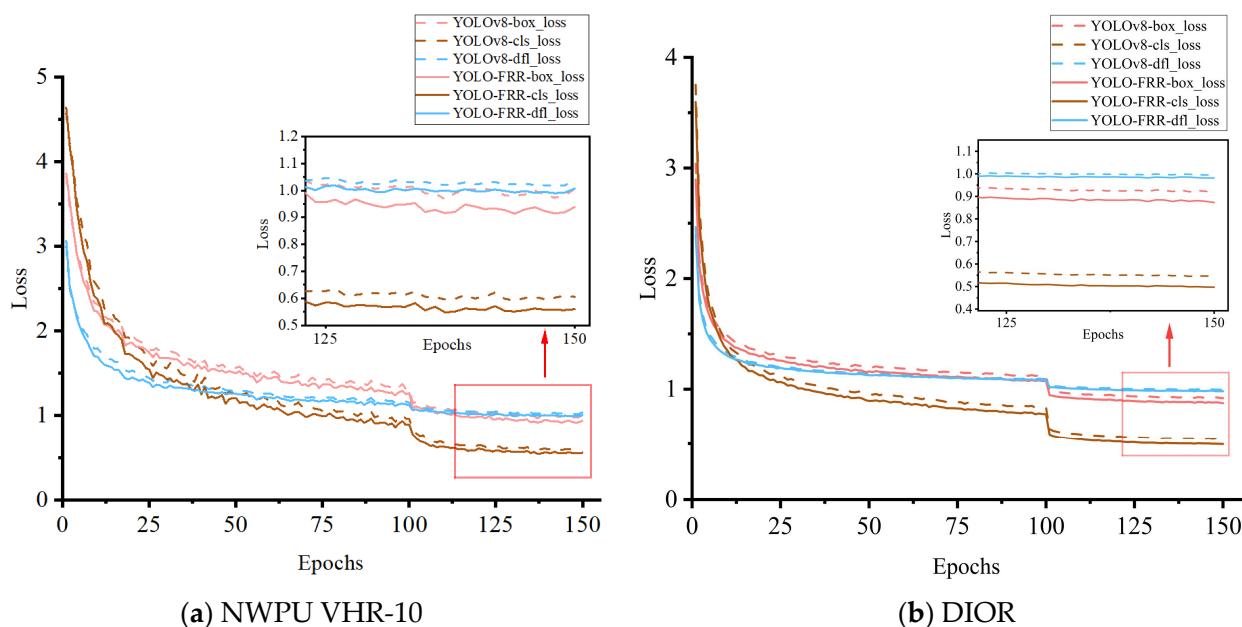


Figure 7. Comparison of training loss between YOLOv8 and FFR-YOLO networks.

Specifically, FFR-YOLO's box_loss decreased faster and stabilized at a lower asymptote, enabling more accurate target spatial feature capture and higher-quality bounding-box regression to reduce baseline localization bias. FFR-YOLO's cls_loss outperformed the baseline consistently, showing enhanced ability to discriminate target classes in complex backgrounds and reduce misclassification. The continued reduction of dfl_loss confirmed FFR-YOLO's advantage in handling boundary uncertainty of remote sensing targets. In the late 100–150 epoch fine-tuning and convergence phase, local enlarged views showed FFR-YOLO's loss continued to decline steadily and remained stable at a low level, while YOLOv8's loss curve flattened and stagnated. This indicated that FFR-YOLO had better optimization stability and convergence performance, and its test set performance improvement further verified its excellent generalization ability.

4.6. Visual Detection Comparison

To evaluate the performance of the proposed FFR-YOLO network architecture in remote sensing object-detection tasks, various test image samples, including scenes with multi-scale objects and densely arranged multi-object scenes, were selected to further analyze the model performance from a visual perspective. Figures 8 and 9 present the visual comparison of detection results between YOLOv8 and FFR-YOLO. The green boxes represent TP, red boxes for FN, and blue boxes FP.

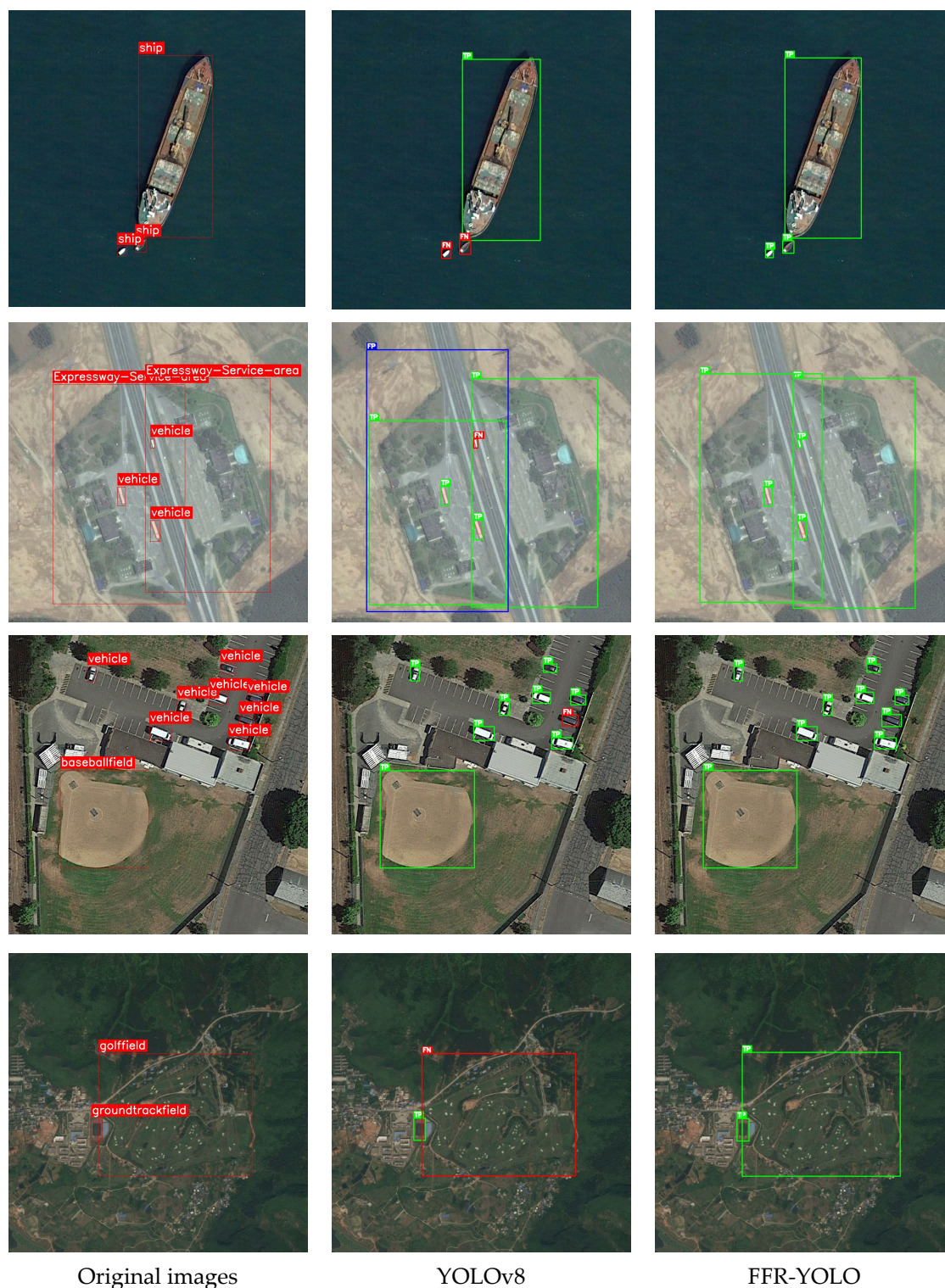


Figure 8. Multi-scale object detection scenes.

Figure 8 illustrates the performance improvement of the FFR-YOLO model in multi-scale object-detection tasks. Multi-scale object distribution maps under four typical scenarios were selected as experimental samples. Objects in these scenarios differed significantly in scale, ranging from extremely small to extremely large sizes. Visual comparison showed that FFR-YOLO achieved superior detection performance in multi-scale object recognition compared with the baseline model YOLOv8. In the illustrated examples, FFR-YOLO delivered superior detection for small-scale objects without sacrificing performance on medium

or large ones. This demonstrated that the proposed model possessed strong adaptability to objects of diverse scales.

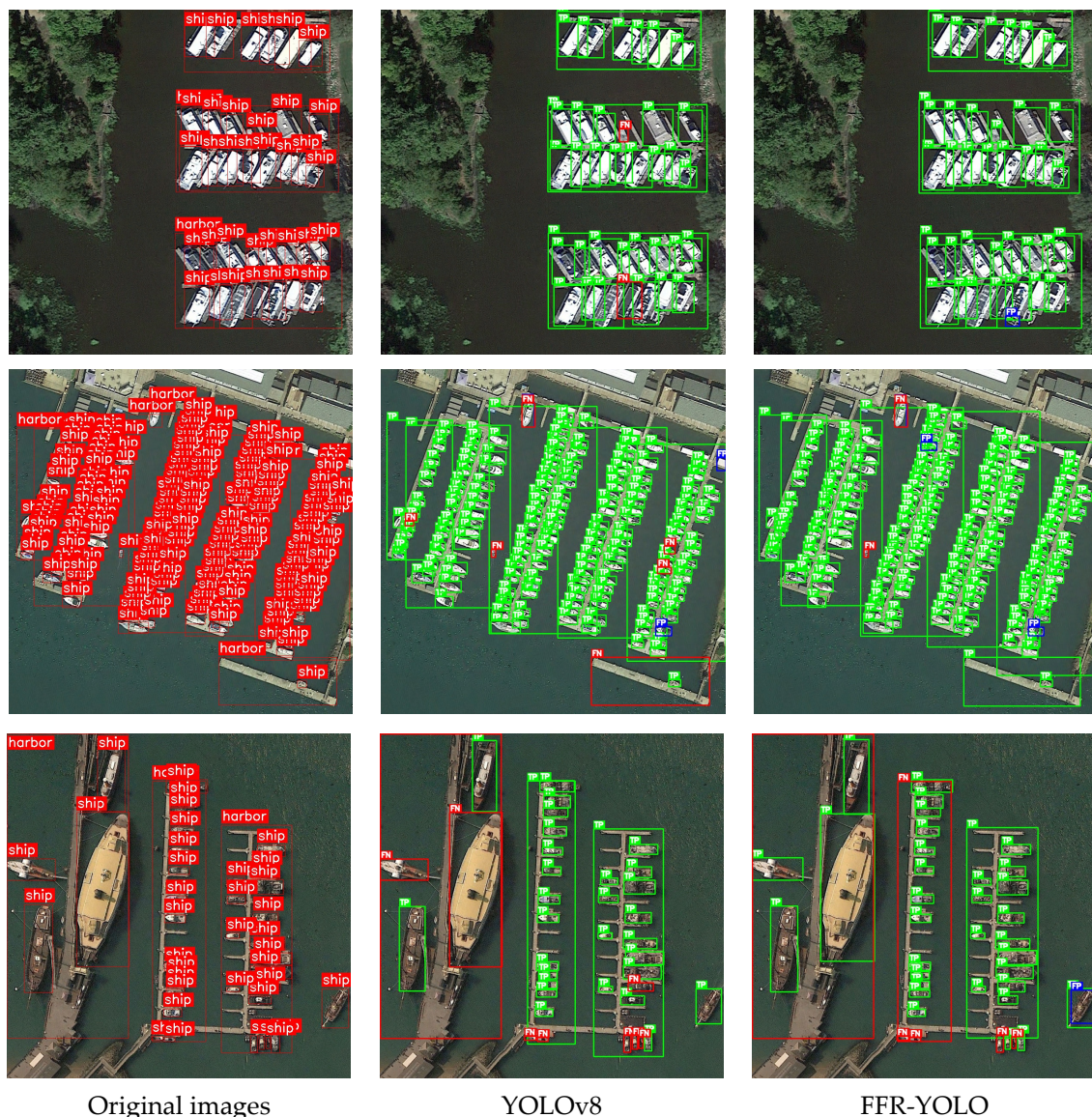


Figure 9. Comparison of detection results for dense objects.

Figure 9 compares and analyzes the detection performance of YOLOv8 and FFR-YOLO in densely arranged remote sensing image object-detection tasks. For dense objects in remote sensing images, FFR-YOLO outperformed YOLOv8 with fewer false detections and missed detections, thus achieving more complete detection results. However, several challenging cases remain. When vessels were packed extremely tightly with severe overlap and minimal inter-object spacing, FFR-YOLO still missed certain instances, particularly the smallest ones whose visual features were nearly indistinguishable from background textures. Additionally, a few false positives appeared in regions where dock structures and moored ships exhibited highly similar edge patterns. Although the proposed method enhanced dense small-object detection performance, the detection of densely packed small objects in crowded scenes remained a direction worth further exploration.

Furthermore, this study compared and analyzed the heatmap results generated by the model before and after network structure optimization when identifying small-sized objects, dense objects, and objects with various scales, as shown in Figures 10–12. These

heatmaps intuitively demonstrated the superiority of FFR-YOLO in reducing false detections and enhancing target response capability in complex scenarios. In the figures, warm-toned regions, such as red and yellow, indicated areas with high model confidence for target existence, highlighting the accurate target localization ability of the model. Such visualization methods facilitated the analysis of the model's attention mechanism towards key regions and helped explain its performance in complex detection tasks.

Taking the scene with dense objects in Figure 10 as an example, FFR-YOLO featured more concentrated response areas, with highlighted regions mostly distributed near the centers of targets and relatively reduced responses to background noise. It proved that the model possessed favorable anti-interference capability and localization performance in complex scenarios. Meanwhile, under conditions of partial occlusion or narrow spacing between objects, FFR-YOLO achieved clearer instance separation, with fewer response adhesion and missed detection cases. This indicated that FFR-YOLO had superior fine-grained representation ability in feature extraction and boundary discrimination, which effectively improved the overall detection quality.

FFR-YOLO achieved prominent improvement in small-object recognition. As shown in Figure 11, the original network presented scattered attention distribution accompanied by massive messy noise in background areas, making small objects easily submerged by complex background interference. By contrast, FFR-YOLO had highly contracted highlighted regions with sharper peaks, which accurately locked the actual positions of aircraft and vehicles and significantly suppressed background noise. The results verified that FFR-YOLO enhanced the focusing ability and localization precision of feature responses and effectively alleviated the common problems of small objects in remote sensing scenarios, such as low pixel proportion and high susceptibility to background disturbance.

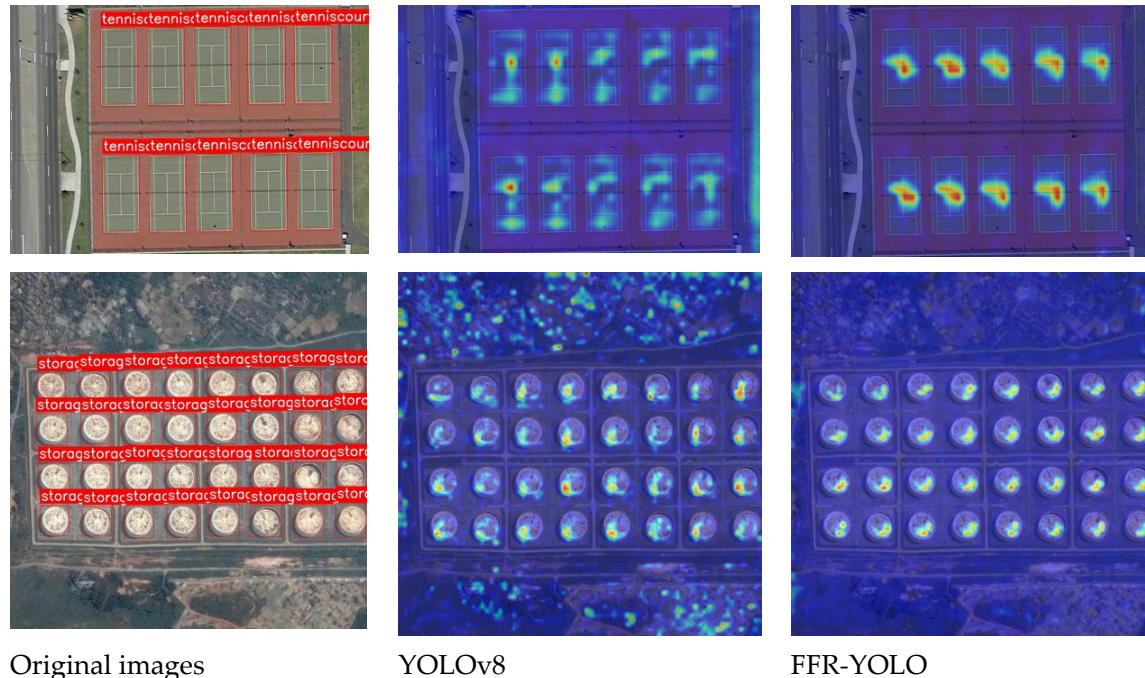


Figure 10. Heatmap of dense object detection.

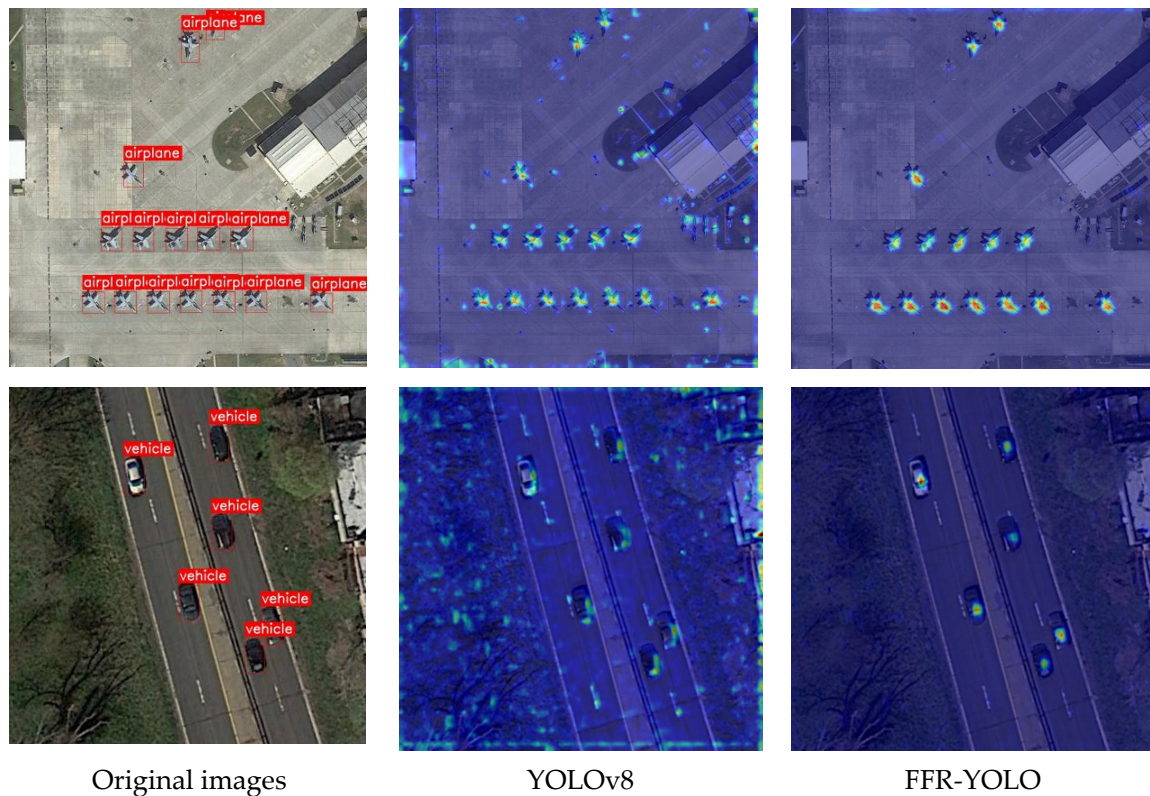


Figure 11. Heatmaps for small object detection.

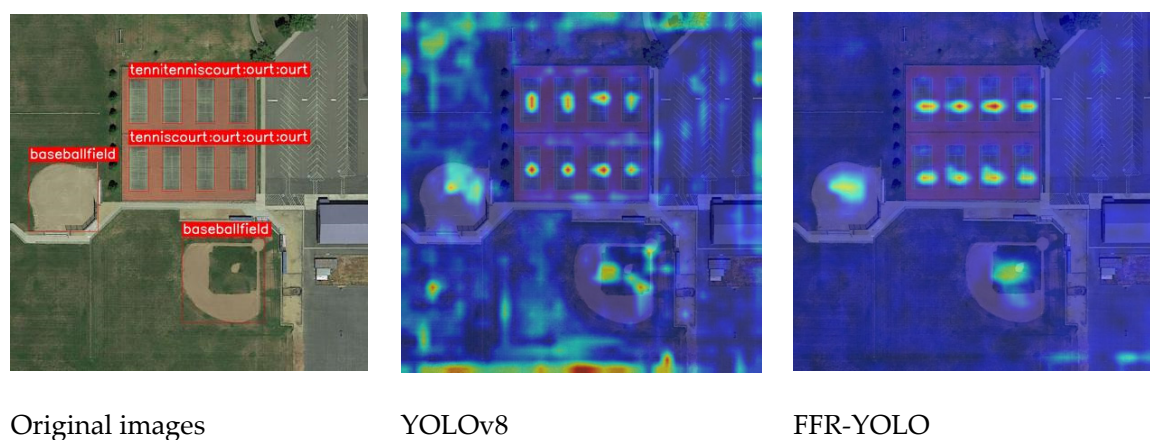


Figure 12. Heatmaps for multi-scale object detection.

For the multi-scale target scenario shown in Figure 12, the heatmap generated by the original network exhibited diffuse and scattered responses for large-scale targets, while the boundaries of medium-scale targets appeared blurred. By contrast, the feature responses generated by the FFR-YOLO network were substantially concentrated. The high-response regions for large-scale targets were precisely focused on the core structural locations of the objects, while background noise fluctuations were significantly suppressed. Concurrently, FFR-YOLO established distinct and clear response regions for medium-sized objects as well. These visual comparisons demonstrated that the proposed FFR-YOLO network could effectively balance feature extraction across diverse scale granularities, thereby maintaining high localization precision even when encountering object combinations with vast dimensional disparity.

4.7. Comparative Analysis of Model Detection Performance

Several classic models in the field of object detection were selected to perform a comparative analysis against the FFR-YOLO model proposed in this study. These included Faster R-CNN, a classic multi-stage detector, along with the classic single-stage detector SSD algorithm and various single-stage detectors from the YOLO series: YOLOv3-tiny, YOLOv5n, YOLOv9t, YOLOv10n, YOLOv11n, YOLOv12n [83], and the latest YOLOv26n [84]. Furthermore, state-of-the-art (SOTA) models proposed by other scholars in recent years, such as Hyper-YOLO and FFCA-YOLO, were included in the comparison. All competing methods were reproduced and trained under the same experimental framework for fair comparison without using original paper results directly. YOLO-series baselines used public versions, all models trained from scratch without pretrained weights. All YOLO-based methods adopted the same training settings to eliminate the influence of different training conditions. Faster R-CNN and SSD also followed the same training protocol to ensure fairness. All architecture-specific hyperparameters were tuned under the same validation protocol and fixed so that performance differences mainly come from detector design rather than unequal training settings. Tables 7 and 8 show the comprehensive comparison of each model's indicators on NWPU VHR-10 and DIOR datasets, respectively.

Table 7. Comparative experiments on NWPU VHR-10 dataset.

	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	Param (M)	FLOPs	FPS
Traditional methods							
Faster R-CNN	44.5	60.0	52.1	31.8	135.9	402.6	6
SSD	57.5	56.8	59.6	37.9	24.1	271.5	18
YOLO Series							
YOLOv3-tiny	95.2	85.2	91.3	57.4	12.1	18.9	105
YOLOv5n	92.5	83.8	90.3	56.6	2.5	7.1	100
YOLOv6	88.6	86.1	90.2	57.6	4.2	11.8	88
YOLOv8n	90.3	83.2	88.3	56.1	3.0	8.1	90
YOLOv9t	93.3	86.4	90.6	57.9	1.9	7.6	65
YOLOv10n	78.9	72.9	81.7	53.8	2.2	6.5	98
YOLOv11n	93.3	85.0	91.8	58.3	2.5	6.3	92
YOLOv12n	91.1	77.5	85.8	58.8	2.5	6.3	81
YOLOv26n	80.2	77.7	84.0	52.1	2.3	5.2	115
SOTA							
Hyper-YOLO	93.9	85.2	92.7	59.4	3.9	10.9	72
FFCA-YOLO	93.3	86.1	93.2	61.4	5.8	15.3	68
Our method							
FFR-YOLO	93.6	88.9	93.6	62.1	4.3	12.6	81

On the NWPU VHR-10 dataset, traditional detectors performed poorly: Faster R-CNN and SSD had mAP50 of 52.1% and 59.6% and inference speeds of 6 and 18 FPS and were unsuited to complex remote sensing scenarios. YOLO-series models had higher accuracy, with YOLOv3-tiny, YOLOv5n, YOLOv9t, and YOLOv11n all exceeding 90% mAP50. The proposed FFR-YOLO reached 93.6% mAP50, 62.1% mAP50:95, and 88.9% recall, outperforming FFCA-YOLO by 0.4 and 0.7 percentage points in mAP50 and mAP50:95. FFR-YOLO inference speed was 81 FPS, higher than Hyper-YOLO's 72 FPS and FFCA-YOLO's 68 FPS and equal to YOLOv12n's 81 FPS, showing a good accuracy–speed trade-off.

On the DIOR dataset with larger class diversity and more complex backgrounds, all models showed degraded performance. FFR-YOLO achieved 79.4% recall, 63.1% mAP50:95, and 85.8% mAP50, outperforming FFCA-YOLO by 0.8, 0.9 and 0.9 percentage points in the three metrics and outperforming Hyper-YOLO by 2.1 and 1.6 percentage points in

recall and mAP50:95. FFR-YOLO ran at 136 FPS, faster than FFCA-YOLO's 123 FPS and Hyper-YOLO's 130 FPS, and remained competitive with lightweight YOLO variants despite higher accuracy. Though YOLOv26n reached 208 FPS, its mAP50 was only 82.5%, while FFR-YOLO balanced detection precision and real-time performance effectively.

Table 8. Comparative experiments on DIOR dataset.

Method	P (%)	R (%)	mAP50 (%)	mAP50:95 (%)	Param (M)	FLOPs	FPS
Traditional methods							
Faster R-CNN	41.8	54.6	47.9	29.7	135.9	402.6	11
SSD	54.2	51.7	56.1	35.4	24.1	271.5	33
YOLO Series							
YOLOv3-tiny	92.4	80.8	87.6	54.2	12.1	18.9	190
YOLOv5n	89.7	79.6	86.9	53.8	2.5	7.1	181
YOLOv6	86.4	81.9	87.2	54.9	4.2	11.8	159
YOLOv8n	86.5	77.7	83.6	60.1	3.0	8.1	163
YOLOv9t	91.2	82.7	88.6	56.4	1.9	7.6	118
YOLOv10n	76.5	70.4	79.8	51.6	2.2	6.5	177
YOLOv11n	91.5	82.1	89.1	57.2	2.5	6.3	166
YOLOv12n	89.6	75.3	84.7	57.5	2.5	6.3	147
YOLOv26n	78.3	75.8	82.5	50.3	2.3	5.2	208
SOTA							
Hyper-YOLO	88.9	77.3	83.7	61.5	3.9	10.9	130
FFCA-YOLO	87.8	78.6	84.9	62.2	5.8	15.3	123
Our Method							
FFR-YOLO	88.5	79.4	85.8	63.1	4.3	12.6	136

4.8. Comparison on Different Datasets

To further validate the effectiveness of our improved remote sensing image recognition method, we conducted generalization experiments using the original YOLOv8 and FFR-YOLO models on three classic remote sensing image datasets: DOTA-v1.5 [85], HRSC2016 [86], and RSOD-Dataset [87]. DOTA-v1.5 is a large-scale benchmark dataset for oriented object detection in aerial images, comprising 16 categories with 403,318 instances. HRSC2016 is a specialized dataset focused on ship detection in high-resolution images, containing 1071 samples with resolutions ranging from 0.5 m to 2 m and varying image sizes. RSOD-Dataset encompasses four object categories, making it suitable for multi-class object detection in the remote sensing domain; it contains a total of 976 images with balanced sample distribution and an average resolution of 0.5 m.

As shown in Table 9, FFR-YOLO consistently outperformed YOLOv8 across all three datasets in terms of mAP while also achieving higher precision and recall, demonstrating its superior generalization capability. Figure 13 presents a visual comparison of the detection results between YOLOv8 and FFR-YOLO on the three datasets. The experimental results demonstrated that FFR-YOLO achieved higher detection accuracy than the baseline network in small-object detection, overlapping-object detection, and multi-scale detection tasks, further validating its strong generalization ability.

Table 9. Comparison results of different datasets.

Datasets	YOLOv8			FFR-YOLO		
	Precision (%)	Recall (%)	mAP50 (%)	Precision (%)	Recall (%)	mAP50 (%)
DOTA-1.5	85.9	81.2	86.7	89.1	83.5	89.9
HRSC-2016	69.2	66.1	68.5	72.5	72.3	74.8
RSOD-Dataset	91.0	86.3	90.5	94.0	87.8	93.2

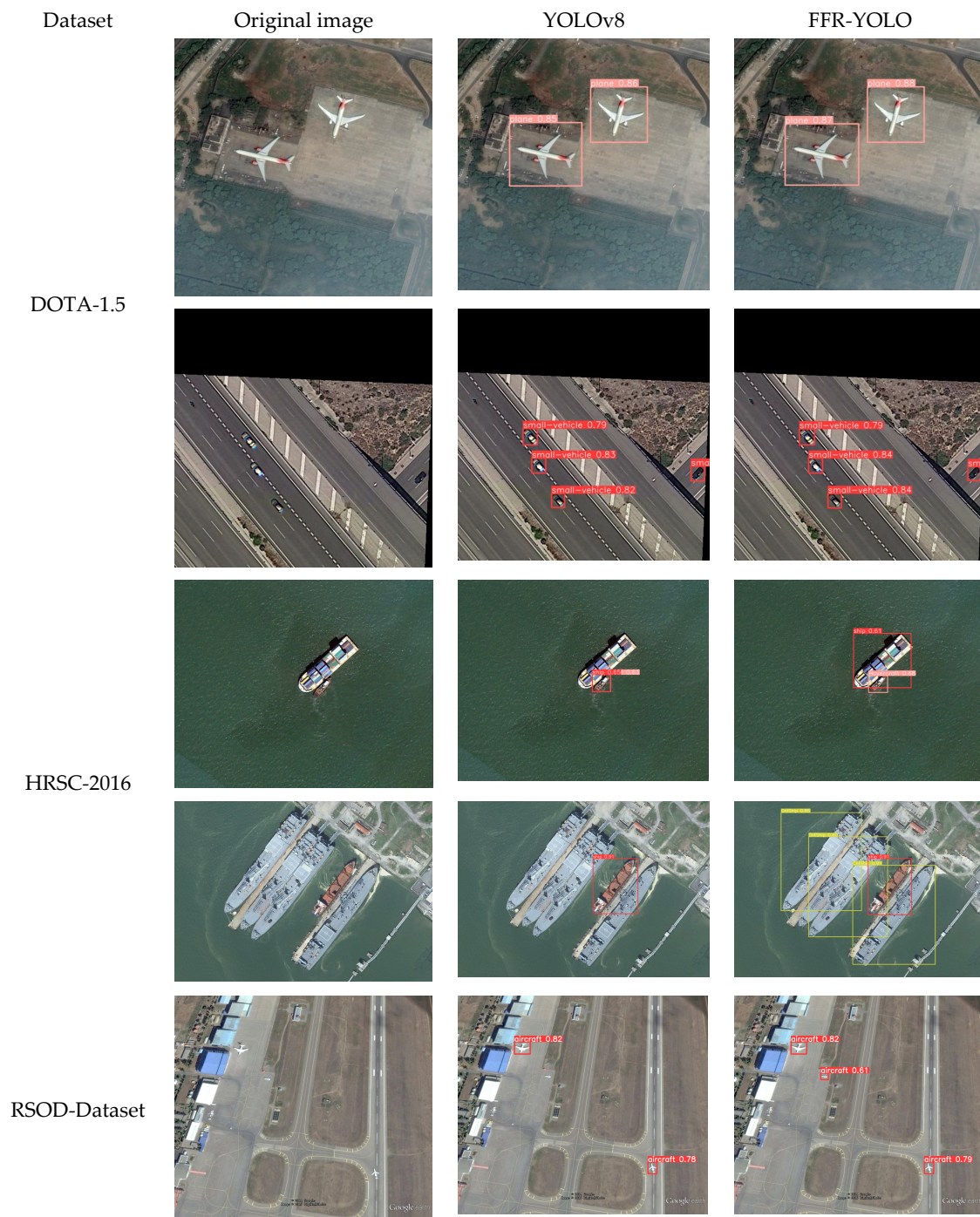


Figure 13. Comparison of images recognition test results on different dataset.

5. Conclusions

To address missed small object detection, difficult dense object separation, and high false detection rates in complex-background remote sensing imagery, this study improved the YOLOv8 detection framework and proposed FFR-YOLO, a frequency-guided fusion reconstruction object detection network. This method performs synergistic optimization at the backbone, neck, and detector head front end: The FAPD module introduced to the backbone uses Haar wavelet decomposition to replace traditional strided convolutional downsampling and combines a low-frequency-guided high-frequency gating mechanism to enhance information preservation; the BRFM designed in the neck strengthens multi-scale feature collaborative expression via multi-source weighted fusion and cross-path bridging

interactions. The RDLGF module constructed at the detector head front end improves the adaptive allocation capability of detection features.

Experimental evaluations on the DIOR and NWPU VHR-10 datasets show this method improves precision, recall, and overall detection performance. Compared with baseline YOLOv8, FFR-YOLO achieves 85.8% mAP50 and 63.1% mAP50:95 on the DIOR dataset and 93.6% and 62.1%, respectively, on the NWPU VHR-10 dataset. The results, including the confusion matrix, verify the method's effectiveness in retaining small object features, separating dense objects, and suppressing complex backgrounds. Ablation and comparative experiments show the FAPD, BRFM, and RDLGF modules function complementarily to achieve an optimized accuracy-efficiency trade-off with minimal increases in parameter and computational overhead, with substantial practical value and application potential. Importantly, the proposed frequency-guided fusion reconstruction strategy is not limited to YOLOv8: FAPD can integrate into any CNN backbone with strided convolutions. BRFM can replace FPN/PAN structures in newer YOLO variants, and RDLGF serves as a general feature recalibration front end for arbitrary detection heads. While adopting more-recent YOLO versions could yield further gains, we selected YOLOv8 for its mature ecosystem and well-established training pipeline, ensuring fair comparisons and reproducible research in the remote sensing community.

Although the method achieved encouraging results, there remains room for improvement. First, the current framework is mainly based on horizontal bounding-box detection, and its adaptability to rotated, elongated targets with strong directionality needs further enhancement. Second, integration of modules such as FAPD introduces additional computational overhead, so subsequent research can adopt lightweight network design to improve deployment feasibility on resource-constrained edge devices. Third, this study's experiments mainly focused on optical remote sensing data; future research can extend to other remote sensing modalities and incorporate advanced techniques to improve the model's generalization and robustness in complex real-world environments.

Author Contributions: Conceptualization, P.Z. and J.Z.; methodology, P.Z.; software, J.Z. and J.L.; validation, J.Z., X.L. and Y.L.; formal analysis, P.Z.; investigation, P.Z.; resources, J.L. and L.T.; data curation, X.L. and Y.L.; writing—original draft preparation, P.Z.; writing—review and editing, J.Z., J.L. and L.T.; visualization, J.Z.; supervision, J.L. and L.T.; project administration, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: China Postdoctoral Science Foundation, the 77th batch of postdoctoral fund surface funding in 2025, funding number: 2025M774430.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: I would like to express my heartfelt gratitude to all the authors whose works have contributed to this paper. Their invaluable insights, rigorous research, and scholarly discussions have greatly enriched my understanding and helped shape the foundation of this study. Without their efforts and dedication to advancing knowledge, this paper would not have been possible.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FFR-YOLO	Frequency-guided Fusion Reconstruction YOLO
FAPD	Frequency-guided Anti-alias Progressive Downsampling Module
BRFM	Bridge-guided Bidirectional Reconstruction Fusion Module
RDLGF	Recalibrated Dual-branch Local–Global Fusion

References

1. Zheng, Z.; Yuan, J.; Yao, W.; Kwan, P.; Yao, H.; Liu, Q.; Guo, L. Fusion of UAV-acquired visible images and multispectral data by applying machine-learning methods in crop classification. *Agronomy* **2024**, *14*, 2670. [\[CrossRef\]](#)
2. Zhang, Z.; Yao, F.; Li, J. Dynamic penetration test based on YOLOv5. In *2022 3rd International Conference on Geology, Mapping and Remote Sensing (ICGMRS)*; IEEE: New York, NY, USA, 2022; pp. 102–105.
3. Morita, M.; Kinjo, H.; Sato, S.; Sulyyon, T.; Anezaki, T. Autonomous flight drone for infrastructure (transmission line) inspection (3). In *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*; IEEE: New York, NY, USA, 2017; pp. 198–201.
4. Menkhoff, T.; Tan, E.K.; Ning, K.S.; Hup, T.G.; Pan, G. Tapping drone technology to acquire 21st century skills: A smart city approach. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*; IEEE: New York, NY, USA, 2017; pp. 1–4.
5. Xu, K.; Song, C.; Xie, Y.; Pan, L.; Gan, X.; Huang, G. RMT-YOLOv9s: An infrared small target detection method based on UAV remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 7002205. [\[CrossRef\]](#)
6. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [\[CrossRef\]](#)
7. Ma, J.; Wu, F.; Li, C.; Tang, C.; Zhang, J.; Xu, Z.; Liu, D. G2EMOT: Guided embedding enhancement for multiple object tracking in complex scenes. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2527214. [\[CrossRef\]](#)
8. Zhou, S.; Liu, Z.; Luo, H.; Qi, G.; Liu, Y.; Zuo, H.; Zhang, J.; Wei, Y. GCA2Net: Global-consolidation and angle-adaptive network for oriented object detection in aerial imagery. *Remote Sens.* **2025**, *17*, 1077. [\[CrossRef\]](#)
9. Liu, X.; Zhou, S.; Ma, J.; Sun, Y.; Zhang, J.; Zuo, H. DFAS-YOLO: Dual feature-aware sampling for small-object detection in remote sensing images. *Remote Sens.* **2025**, *17*, 3476. [\[CrossRef\]](#)
10. Liu, Z.; Chen, C.; Huang, Z.; Chang, Y.C.; Liu, L.; Pei, Q. A low-cost and lightweight real-time object-detection method based on uav remote sensing in transportation systems. *Remote Sens.* **2024**, *16*, 3712. [\[CrossRef\]](#)
11. Navneet, D. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*; IEEE: New York, NY, USA, 2005; Volume 2, pp. 886–893.
12. Lowe, D.G. Object recognition from local scale-invariant features. In *Seventh IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 1999; Volume 2, pp. 1150–1157.
13. Pan, W.; Huan, W.; Xu, L. Improving High-Voltage Line Obstacle Detection with Multi-Scale Feature Fusion in YOLO Algorithm. In *2024 6th International Conference on Electronics and Communication, Network and Computer Technology (ECNCT)*; IEEE: New York, NY, USA, 2024; pp. 270–273.
14. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [\[CrossRef\]](#)
15. Wang, Z.; Xia, F.; Zhang, C. FD_YOLOX: An improved YOLOX object detection algorithm based on dilated convolution. In *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*; IEEE: New York, NY, USA, 2023; pp. 1263–1268.
16. Li, H. Rethinking features-fused-pyramid-neck for object detection. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2025; pp. 74–90.
17. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. CGNet: A light-weight context guided network for semantic segmentation. *IEEE Trans. Image Process.* **2020**, *30*, 1169–1179. [\[CrossRef\]](#)
18. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [\[CrossRef\]](#)
19. Li, C.; Zhao, R.; Wang, Z.; Xu, H.; Zhu, X. Remdet: Rethinking efficient model design for uav object detection. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 4643–4651. [\[CrossRef\]](#)
20. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.
21. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In *IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 5449–5457.
22. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 10781–10790.
23. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic detr: End-to-end object detection with dynamic attention. In *IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 2988–2997.
24. Wang, X.; He, N.; Hong, C.; Wang, Q.; Chen, M. Improved YOLOX-X based UAV aerial photography object detection algorithm. *Image Vis. Comput.* **2023**, *135*, 104697. [\[CrossRef\]](#)
25. Shi, Y.; Wang, C.; Xu, S.; Yuan, M.D.; Liu, F.; Zhang, L. Deformable convolution-guided multiscale feature learning and fusion for UAV object detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6004105. [\[CrossRef\]](#)
26. Zhuang, Z.; Wang, Z.; Chen, S.; Liu, L.; Luo, H.; Tan, M. Robust 3d semantic occupancy prediction with calibration-free spatial transformation. *arXiv* **2024**, arXiv:2411.12177.

27. Debnath, D.; Vanegas, F.; Sandino, J.; Hawary, A.F.; Gonzalez, F. A review of UAV path-planning algorithms and obstacle avoidance methods for remote sensing applications. *Remote Sens.* **2024**, *16*, 4019. [\[CrossRef\]](#)
28. Xue, W.; Ai, J.; Zhu, Y.; Sun, X.; Zhang, Y.; Gao, G. LMCNet: Lightweight Modality Compensation Network via Knowledge Distillation for Salient Ship Detection under Missing-Modality Conditions. *IEEE Trans. Aerosp. Electron. Syst.* **2026**, *62*, 6547–6560. [\[CrossRef\]](#)
29. Ai, J.; Tian, R.; Luo, Q.; Jin, J.; Tang, B. Multi-Scale Rotation-Invariant Haar-Like Feature Integrated CNN-Based Ship Detection Algorithm of Multiple-Target Environment in SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10070–10087. [\[CrossRef\]](#)
30. Chen, E.A. SAR Target Classification Using the Multikernel-Size Feature Fusion-Based Convolutional Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5214313.
31. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*; IEEE: New York, NY, USA, 2001; Volume 1, p. I.
32. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote Sensing Object Detection in the Deep Learning Era—A Review. *Remote Sens.* **2024**, *16*, 327. [\[CrossRef\]](#)
33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2014; pp. 580–587.
34. Girshick, R. Fast r-cnn. In *IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
35. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 91–99. [\[CrossRef\]](#)
36. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 6154–6162.
37. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
38. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 4700–4708.
39. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 779–788.
40. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2017; pp. 7263–7271.
41. Redmon, J. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
42. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolo4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
43. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; NanoCode012; Kwon, Y.; TaoXie; Michael, K.; Fang, J.; imyhxy; et al. YOLOv5. 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 1 May 2026).
44. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
45. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 7464–7475.
46. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 May 2026).
47. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolo9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 1–21.
48. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolo10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [\[CrossRef\]](#)
49. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
50. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
51. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
52. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. Detrs beat yolos on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 16965–16974.
53. Wang, K.; Wang, Z.; Li, Z.; Su, A.; Teng, X.; Pan, E.; Liu, M.; Yu, Q. Oriented object detection in optical remote sensing images using deep learning: A survey. *Artif. Intell. Rev.* **2025**, *58*, 350. [\[CrossRef\]](#)
54. Koyun, O.C.; Keser, R.K.; Akkaya, I.B.; Töreyin, B.U. Focus-and-Detect: A small object detection framework for aerial images. *Signal Process. Image Commun.* **2022**, *104*, 116675. [\[CrossRef\]](#)

55. Lin, Y.; Li, J.; Shen, S.; Wang, H.; Zhou, H. GDRS-YOLO: More Efficient Multiscale Features Fusion Object Detector for Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6008505. [\[CrossRef\]](#)
56. Li, M.; Chen, Y.; Zhang, T.; Huang, W. TA-YOLO: A Lightweight Small Object Detection Model Based on Multi-Dimensional Trans-Attention Module for Remote Sensing Images. *Complex Intell. Syst.* **2024**, *10*, 5459–5473. [\[CrossRef\]](#)
57. Yang, G.; Lei, J.; Tian, H.; Feng, Z.; Liang, R. Asymptotic feature pyramid network for labeling pixels and regions. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7820–7829. [\[CrossRef\]](#)
58. Zhang, F.; Tian, C.; Li, X.; Yang, N.; Zhang, Y. AL-YOLOv8: A Small Object Detection Algorithm for Remote Sensing Images Based on an Improved YOLOv8s. *Sensors* **2026**, *26*, 2016. [\[CrossRef\]](#)
59. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe++: Unified content-aware reassembly of features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4674–4687. [\[CrossRef\]](#)
60. Chen, Z.; Ma, Y.; Gong, Z.A.; Cao, M.; Yang, Y.; Wang, Z.; Wang, T.; Li, J.; Liu, Y. R-AFPN: A residual asymptotic feature pyramid network for UAV aerial photography of small targets. *Sci. Rep.* **2025**, *15*, 16233. [\[CrossRef\]](#)
61. Aghili, M.E.; Ghassemian, H.; Imani, M. YOLO-PICO: Lightweight Object Recognition in Remote Sensing Images using Expansion Attention Modules. *Pattern Recognit.* **2026**, *176*, 113114. [\[CrossRef\]](#)
62. Wu, Y.; Mu, X.; Shi, H.; Hou, M. An object detection model AAPW-YOLO for UAV remote sensing images based on adaptive convolution and reconstructed feature fusion. *Sci. Rep.* **2025**, *15*, 16214. [\[CrossRef\]](#)
63. Meng, L.; Li, D.; He, J.; Ma, L.; Li, Z. Convolutional Feature Enhancement and Attention Fusion BiFPN for Ship Detection in SAR Images. *arXiv* **2025**, arXiv:2506.15231.
64. Mao, Y.; Zhang, H.; Li, R.; Zhu, F.; Sun, R.; Ji, P. HSF-DETR: Hyper scale fusion detection transformer for multi-perspective UAV object detection. *Remote Sens.* **2025**, *17*, 1997. [\[CrossRef\]](#)
65. Liu, J.; Tao, J.; Liu, X.; Ma, J.; Guo, C.; Dong, C.; Li, P.; Shi, P. Multi path attention and scale aware fusion for accurate object detection in remote sensing imagery. *Sci. Rep.* **2025**, *15*, 41810. [\[CrossRef\]](#)
66. Hu, Z.; Chen, W.; Yang, D.; Wang, Z.; Wang, P. Enhancing object detection in remote sensing images with improved YOLOv8 model. *Sci. Rep.* **2025**, *15*, 42488. [\[CrossRef\]](#)
67. Shi, J.; Liu, J.; Zhang, J.; Zhang, L.; Sun, S. OA-YOLOv8: A Multiscale Feature Optimization Network for Remote Sensing Object Detection. *Appl. Sci.* **2026**, *16*, 1467. [\[CrossRef\]](#)
68. Luo, X.; Peng, Y.; Xie, R.; Li, P.; Qian, Y. Enhanced Multi-Scale Feature Extraction Lightweight Network for Remote Sensing Object Detection. *Comput. Mater. Contin.* **2026**, *86*, 90. [\[CrossRef\]](#)
69. Wang, X.; Peng, L.; Li, X.; He, Y. Enhancing remote sensing object detection via selective-perspective-class integration. *Eng. Appl. Artif. Intell.* **2026**, *165*, 113416. [\[CrossRef\]](#)
70. Chen, M.; Wang, C.; Yu, Y.; Li, Y.; Li, Y.; Chen, X. Improved YOLOv9-based remote sensing image detection method. *Sci. Rep.* **2025**, *15*, 44721. [\[CrossRef\]](#)
71. Bao, W.; Wang, P.; Hu, W. MSPA-YOLO: Remote Sensing Target Recognition Method Based on Improved YOLOv10. In *2025 IEEE International Conference on Pattern Recognition, Machine Vision and Artificial Intelligence (PRMVAI)*; IEEE: New York, NY, USA, 2025; pp. 1–5.
72. Anandakrishnan, J.; Sangaiah, A.K.; Darmawan, H.; Son, N.K.; Lin, Y.B.; Alenazi, M.J. Precise spatial prediction of rice seedlings from large-scale airborne remote sensing data using optimized Li-YOLOv9. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *18*, 2226–2238. [\[CrossRef\]](#)
73. Li, B.; Kang, Y.; Ding, Y.; Li, S.; Zhang, Z.; Ma, D. DAE-YOLO: Remote Sensing Small Object Detection Method Integrating YOLO and State Space Models. *Remote Sens.* **2025**, *18*, 109. [\[CrossRef\]](#)
74. Ji, C.L.; Yu, T.; Gao, P.; Wang, F.; Yuan, R.Y. YOLO-TLA: An efficient and lightweight small object detection model based on YOLOv5. *J. Real-Time Image Process.* **2024**, *21*, 141. [\[CrossRef\]](#)
75. Liao, H.; Zhu, W. YOLO-DRS: A bioinspired object detection algorithm for remote sensing images incorporating a multi-scale efficient lightweight attention mechanism. *Biomimetics* **2023**, *8*, 458. [\[CrossRef\]](#)
76. Chen, Y.; Yuan, X.; Wang, J.; Wu, R.; Li, X.; Hou, Q.; Cheng, M.M. YOLO-MS: Rethinking multi-scale representation learning for real-time object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4240–4252. [\[CrossRef\]](#)
77. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [\[CrossRef\]](#)
78. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-Class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [\[CrossRef\]](#)
79. Rahman, M.M.; Munir, M.; Marculescu, R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2024; pp. 11769–11779.
80. Zhang, J.; Zhang, Y.; Shi, Z.; Zhang, Y.; Gao, R. Unmanned aerial vehicle object detection based on information-preserving and fine-grained feature aggregation. *Remote Sens.* **2024**, *16*, 2590. [\[CrossRef\]](#)

81. Qi, Y.; Cao, H. SDFA-Net: Synergistic Dynamic Fusion Architecture With Deformable Attention for UAV Small Target Detection. *IEEE Access* **2025**, *13*, 110636–110647. [[CrossRef](#)]
82. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
83. Tian, Y.; Ye, Q.; Doermann, D. Yolov12: Attention-centric real-time object detectors. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 78433–78457.
84. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv* **2025**, arXiv:2509.25164.
85. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: New York, NY, USA; pp. 3974–3983.
86. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 24–26 February 2017*; SciTePress: Setúbal, Portugal, 2017; Volume 2017, pp. 324–331.
87. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.