ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 917-928
DOI 10.5755/j01.itc.55.2.44290

DHAF-YOLO: Dynamic Hierarchical Attention Fusion for Small Object Detection

Received 2026/02/12 | Accepted after revision 2026/06/03

HOW TO CITE: Wang, Z., Wang, X., Yao, S. (2026) DHAF-YOLO: Dynamic Hierarchical Attention Fusion for Small Object Detection. *Information Technology and Control*, 55(2), 917-928. https://doi.org/10.5755/j01.itc.55.2.44290

DHAF-YOLO: Dynamic Hierarchical Attention Fusion for Small Object Detection

Zhiyue Wang #
School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, 150001, China

Xingjie Wang #
Department of Natural Sciences, Durham University, Durham City, DH1 5TS, United Kingdom

Siyu Yao*
Ocean University of China, Haide College, Qingdao, 266100, China

# Zhiyue Wang and Xingjie Wang contributed equally to this work.

Corresponding author: yaosiyu@stu.ouc.edu.cn

Small object detection in remote-sensing images remains challenging due to weak object features, scale variation, and complex background interference. To address these issues, this paper proposes DHAF-YOLO, a dynamic hierarchical attention fusion detector based on YOLOv11. The proposed method integrates three key designs: backbone enhancement, hierarchical cross-scale fusion, and adaptive focus refinement. First, Dynamic Weighted Feature Convolution (DWFC) is introduced to generate spatially sensitive responses and enhance fine-grained feature extraction. Second, Cross-Scale Gated Fusion (CSGF), implemented through dynamic scale-sequence fusion, performs two-stage interaction across P2–P5 pyramid levels to preserve high-resolution details and improve semantic consistency. Third, Dual-Path Adaptive Focus (DPAF) further refines fused features to strengthen target-related responses and suppress background noise. Experiments on the VEDAI-Coco dataset show that DHAF-YOLO achieves the highest Recall, mAP@0.5, and mAP@0.5:0.95 among compared YOLO-series detectors from YOLOv5 to YOLOv12, although its Precision is lower than YOLOv9. Compared with YOLOv9, DHAF-YOLO improves Recall by 2.4 percentage points and mAP@0.5:0.95 by 5.0 percentage points, while Precision decreases by 4.0 percentage points. These results demonstrate its effectiveness for remote-sensing small object detection.

KEYWORDS: Small Object Detection, Remote Sensing Image Analysis, YOLO-based Detector, Hierarchical Attention Fusion, Multi-scale Feature Learning

1. Introduction

Advancements in high-resolution optical remote sensing have greatly promoted the application of object detection in aerial and spaceborne imagery. These techniques are increasingly important in environmental surveillance, disaster response, urban development, and defence reconnaissance [12]. A remote-sensing image usually covers a large geographical area and contains objects distributed at very different scales. However, due to acquisition conditions and the physical limitations of imaging sensors, many targets suffer from low imaging quality, especially small objects whose size is smaller than 32 × 32 pixels. Such objects contain only limited visual cues, making it difficult to describe their semantics and causing object detection to remain a challenging task [6, 23, 29].

At the same time, uncontrolled environmental conditions, including platform motion, weather variation, atmospheric turbulence, and scene diversity, often lead to strong aliasing between objects and the background, which further increases the difficulty of detecting small objects in remote-sensing images [18]. This challenge is particularly pronounced in remote-sensing scenarios, where small targets are easily submerged in cluttered background patterns and therefore become difficult to localize and recognize accurately [9]. Consequently, enhancing feature discrimination under complex backgrounds is of great importance for improving the robustness and reliability of small object detection [1].

The primary obstacles facing small object detection in remote-sensing imagery can be categorized into three main aspects. First, feature representation is often insufficient, since small targets possess limited visual cues and tend to disappear during deep convolutional operations [15]. Second, background interference is severe, because complex environments frequently blur the boundaries between objects and their surroundings [20]. Third, scale diversity and feature fragmentation remain difficult to handle, as the varied size distribution of small targets often causes local details to be lost under fixed-scale feature extraction [14].

To address these issues, this work aims to develop an accurate remote-sensing small object detector based on dynamic feature enhancement and hierarchical attention fusion. Specifically, a Dynamic Weighted Feature Convolution (DWFC) module is proposed to generate spatially adaptive responses and improve multi-scale perception for weak targets. A Cross-Scale Gated Fusion (CSGF) module is then introduced to enhance multi-scale feature integration through explicit feature alignment and cross-scale interaction. Finally, a contextual refinement strategy based on hierarchical fusion and a Dual-Path Adaptive Focus (DPAF) module is employed to model contextual dependencies and suppress irrelevant background information. By integrating these components into YOLOv11, we build a new detection framework named Dynamic Hierarchical Attention Fusion YOLO (DHAF-YOLO).

To cope with the scale variations of small objects in remote sensing images, a scale-adaptive component is incorporated into the module architecture. It allows the model to dynamically tune the feature extraction granularity, thereby accommodating the varied size distributions of these targets and preventing the information loss often associated with static-scale methods.

The main contributions of this study are summarized as follows:

1. A remote-sensing small-object detector, termed DHAF-YOLO, is developed based on YOLOv11 and achieves strong detection performance on the VEDAI-Coco dataset.
2. A dynamic backbone enhancement strategy based on DWFC is introduced to improve local feature extraction and fine-grained target representation.
3. A hierarchical fusion and adaptive refinement pipeline is designed to strengthen cross-scale interaction, preserve shallow spatial details, and suppress background interference in complex remote-sensing scenes.

The rest of this article is organized as follows. Section 2 reviews prior research on small object detection, including YOLO-based methods in remote sensing, multi-scale feature fusion, and contextual modeling. Section 3 describes the proposed DHAF-YOLO detector in detail. Section 4 presents the experimental settings, ablation studies, and comparative results. Finally, Section 5 concludes the paper and discusses possible future research directions.

2. Related Works

This section reviews previous studies related to the proposed method. In particular, we summarize representative works on YOLO-based detection in remote sensing, multi-scale feature fusion, global context modeling, and scale-adaptive feature processing.

2.1. YOLO-based Methods in Remote Sensing

Deep learning-based object detectors have revolutionized remote sensing analysis by enabling end-to-end feature extraction and object localisation [32]. Existing detection systems typically fall into two primary classes: two-stage and one-stage architectures. Although there is a moderate trade-off in accuracy, one-stage models offer significantly reduced computational costs compared to two-stage methods, making them ideal for time-sensitive scenarios [27, 33]. Notably, the YOLO series has proven effective in handling aerial and spaceborne data, demonstrating superior scalability and efficiency among one-stage detectors.

Several improvements to YOLO have been proposed for remote sensing applications, including the incorporation of transformer encoder blocks to enrich global dependencies and deformable convolutions for adaptive feature alignment. For instance, attention-based YOLO variants can suppress background noise and enhance representation through long-range pixel dependencies [13, 24]. While these modifications yield accuracy gains, they sometimes introduce considerable computational overhead, which may limit their practicality in resource-constrained application scenarios. Given the favourable trade-off between efficiency and accuracy, YOLO is selected as the baseline architecture in this study, upon which task-specific modules are incorporated to enhance small-object feature representation and suppress background interference [25].

2.2. Enhancement and Fusion of Multi-Scale Features

In remote sensing tasks, small targets often occupy minimal spatial area and may correspond to only a few pixels in deep feature maps. It necessitates a robust, multi-scale representation for effective discrimination [28]. Multi-scale fusion methods, inspired by feature pyramids, combine high-resolution shallow features with semantically rich deep features to bridge the scale gap[7, 8]. Enhanced fusion mechanisms, such as adaptive fusion weights or attention to salient regions can further improve discrimination. However, many existing designs rely on static weighting or heuristic alignment, which limits adaptability to rapidly changing backgrounds and scales [30, 31].

In this work, the DWFC module is introduced for dynamic multi-scale enhancement, expanding receptive fields and adaptively weighting spatial regions without substantial parameter growth. Additionally, the CSGF module efficiently re-weights and aligns multi-scale features, achieving superior fusion performance with minimal resource consumption.

2.3. Global Context Modeling

After enhancing local and multi-scale features, modelling global dependencies between targets and scene context further improves the localization of small objects. Non-local attention and context aggregation networks demonstrate that establishing global pixel associations benefits detection accuracy. However, many conventional approaches suffer from high computation or fuse irrelevant background cues [10, 19, 22].

Inspired by efficient context aggregation strategies, our Dual-Path Adaptive Focus (DPAF) module employs both global average pooling and spatial attention to guide cross-channel and spatial interactions, selectively strengthening object-specific features while suppressing distractors.

2.4. Scale-Adaptive Feature Processing for Small Targets

Multi-scale adaptability is crucial for detecting small objects in remote sensing, as objects of various sizes often coexist within the same image [11]. Traditional methods using fixed-scale features to extract information on small objects are difficult to extract complete and stable information, particularly when coupled with the characteristics of multi-scale targets, resulting in false detection and missed detection [21].

Existing attempts at scale adaptation mainly rely on pre-set multi-scale anchors or static pyramid adjustments, which lack flexibility in dynamic scenes. For example, some methods use manually designed anchor boxes for specific scales, but they fail to adapt to unexpected scale variations in complex remote

sensing scenarios[3, 5, 17, 26]. Other works adopt adaptive pyramid levels, but they still rely on fixed interval adjustments and cannot dynamically match the real-time scale of small targets [2, 4, 16, 34].

In this paper, we focus on the research of dynamic-scale feature extraction. By adaptively adjusting the granularity of feature extraction in response to predicted object scales, the model can flexibly extract features of appropriate granularity for small objects, thereby mitigating scale conflicts across different resolution levels.

3. Proposed Method

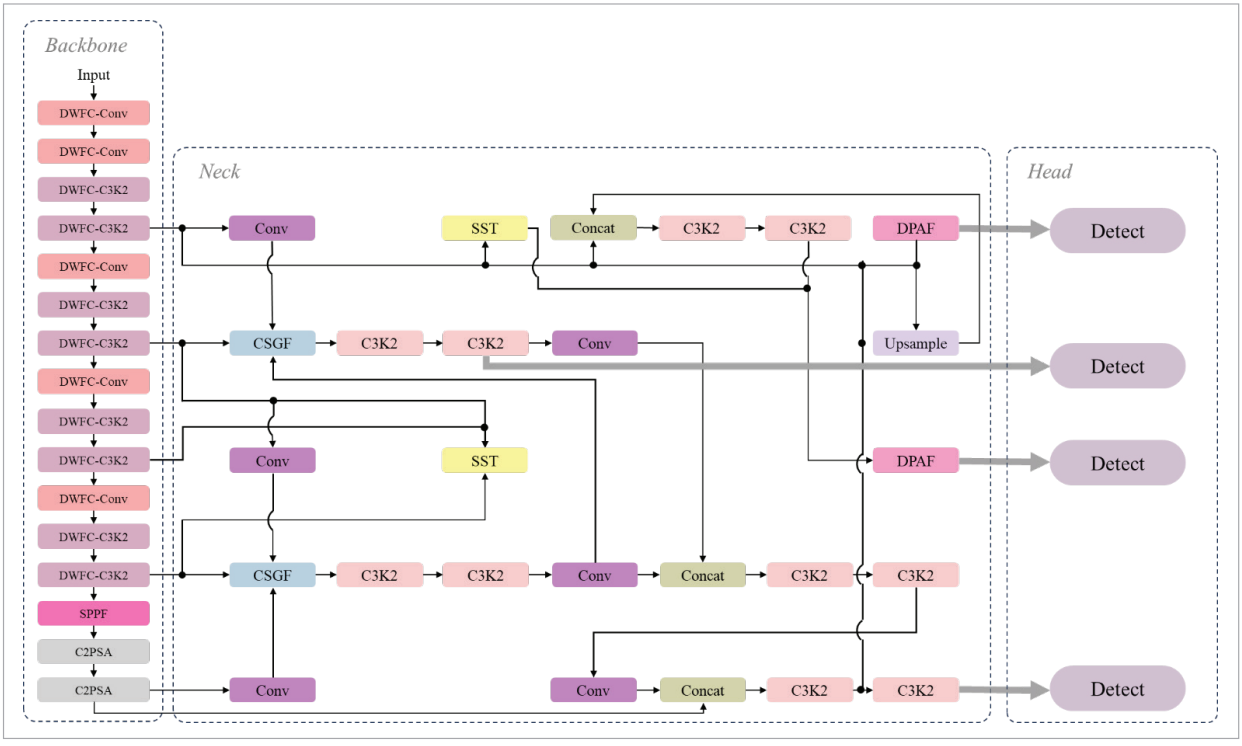
3.1. Overview

We adopted YOLOv11 as the basic framework because of its favorable trade-off between accuracy and efficiency. To address the challenges of weak small-target features, large scale variation, and complex background interference, we propose DHAF-YOLO, a dynamic hierarchical attention fusion detector for remote-sensing images. The proposed detector consists of three main architectural enhancements. First, in the backbone, a Dynamic Weighted Feature Convolution (DWFC) module is employed to improve local response sensitivity and multi-scale feature representation. In the implementation, DWFC is realized by RFCAConv and C3k2_RFCAConv blocks. Second, in the neck and head, a hierarchical fusion pipeline is introduced to align and aggregate pyramid features from different resolutions; its core cross-scale interaction is implemented by DynamicScalSeq. Third, a Dual-Path Adaptive Focus (DPAF) module, implemented by asf_attention_model, is applied after cross-scale fusion to suppress background noise and enhance target-related responses.

As depicted in Figure 1, the backbone first extracts hierarchical feature maps at four pyramid levels, namely P2, P3, P4, and P5. The subsequent feature

Figure 1

Overall framework of DHAF-YOLO. The detector adopts a four-level pyramid structure, namely P2, P3, P4, and P5. Cross-scale fusion is performed in two stages: P3–P5 are first fused to enhance semantic consistency, and P2 is then incorporated to preserve high-resolution details for tiny-object detection. The final detection layer receives four refined features corresponding to P2, P3, P4, and P5.



interaction is conducted in two stages. In the first stage, the higher-level pyramid features P3, P4, and P5 are fused to strengthen semantic consistency across middle and deep feature levels. In the second stage, the shallow P2 feature is further incorporated into the refined representation to preserve high-resolution spatial details that are particularly important for tiny-object detection. After adaptive focus refinement, the final detector outputs predictions from four detection branches associated with P2, P3, P4, and P5, respectively.

3.2. Dynamic Weighted Feature Convolution (DWFC)

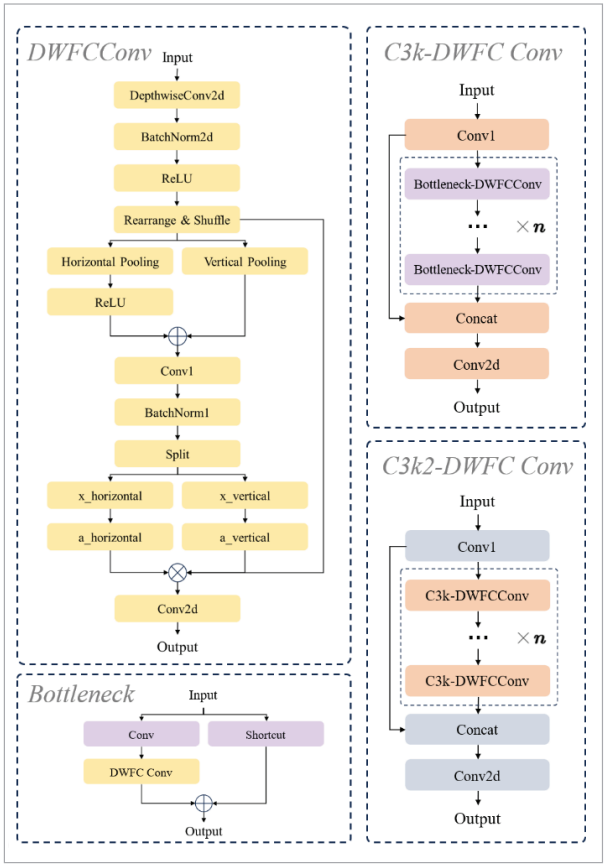
The proposed Dynamic Weighted Feature Convolution (DWFC) module is designed to improve local feature extraction for small objects by combining dynamic response generation and adaptive spatial reweighting. As shown in Figure 2, DWFCConv is used as the core operator, and a Bottleneck structure with DWFCConv is further built to strengthen feature extraction. In particular, the proposed C3k2-DWFC structure enhances deep feature stacking and multi-scale perception, allowing the network to focus more effectively on fine-grained target cues hidden in complicated backgrounds.

After the input data is first processed by the DWFC-Conv operator, which directly associates data from multiple spatial points by performing a deep convolution in the spatial dimension, the output passes through a batch normalization layer and a ReLU activation function. Next is the reshuffle operation, which enables the spatial-wise feature structure to capture more neighbouring pixels based on the channel shuffle. Of the two branches of the Strip Pooling module, a global average pooling operation is applied in each branch to encode two attention maps, which carry long-range dependencies in the horizontal and vertical directions, respectively. The concatenated two descriptors are fused by a 1×1 convolution, BN and ReLU operation. Then, they are split, and two 1×1 convolution and sigmoid operations are applied to each to generate two spatial attention maps, ah and aw, for the horizontal and vertical stripes, respectively. The attention maps are then used to recalibrate the input feature map adaptively through multiplication. A regular convolution is finally used to create the output.

We embed the DWFCConv into a basic network block, named Bottleneck with DWFCConv, where the DWFCConv is added on top of a standard 1×1 Conv layer. The residual architecture mitigates degradation and facilitates the backward propagation of training in deep networks. According to the bottleneck structure, the C3k2-DWFC module first uses a 1×1 convolution to reduce channel dimensions. Then, *n* C3k-DWFCConv modules are stacked to extract rich features, where each C3k-DWFCConv combines *n* independent Bottleneck with DWFCConv modules, to detect scale-aware features on a relatively deep 2D network. The C3k2-DWFC module finally employs a 1×1 convolution to recover the original channel number by aggregating the concatenation of inputs with newly extracted features, thereby maintaining as much information as possible.

Figure 2

Structure of DWFC model global contextual relationships, and spatial saliency across multiple scales.



Let us now go deeper into the processing performed by the DWFCConv:

$$F_{trans}=R(ReLU(BatchNorm(DepthwiseConv2d(X)))) \quad (1)$$

$$F_h=GAPh(F_{trans}),\ F_w=GAPw(F_{trans}) \quad (2)$$

$$Y=ReLU(BatchNorm(Conv1\times1(Concat(Fh,Fw)))) \quad (3)$$

$$a_h=\sigma(Conv_{1\times1}(Y_h)),\ a_w=\sigma(Conv_{1\times1}(Y_w)) \quad (4)$$

$$F_{weighted}=a_h\odot a_w\odot F_{trans} \quad (5)$$

$$Output=Conv_{std}(F_{weighted}), \quad (6)$$

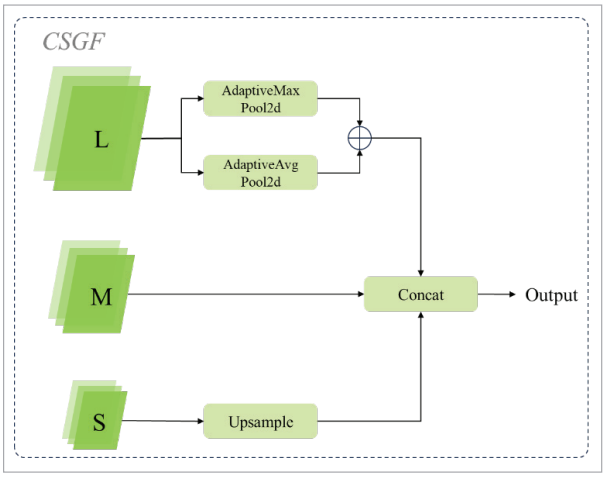
where X denotes the input feature, R denotes the reshape operation, GAP_h and GAP_w represent global average pooling along the height and width dimensions, respectively. The symbol σ denotes the sigmoid activation function, $\odot$ denotes element-wise multiplication, and $Conv_{std}$ denotes the standard convolution after adaptive recalibration.

3.3. Multi-scale Feature Fusion Module

To better utilize multi-level features for remote-sensing small-object detection, the proposed multi-scale fusion design adopts a two-stage cross-scale interaction strategy. In the first stage, the P3, P4, and P5 features are fused to improve semantic consistency among middle and deep pyramid levels. In the second stage, the shallow P2 feature is further introduced together with the refined higher-level features, so that high-resolution localization cues can be preserved for tiny-object detection.

In the implementation, CSGF is realized by the DynamicScalSeq module. Specifically, multiple pyramid features are first projected to compatible channel dimensions and spatially aligned. Then, scale-sequence aggregation is performed to exchange information among different pyramid levels. Unlike a purely point-wise projection, this operation is designed to aggregate multi-level responses and generate a scale-aware representation for subsequent detection. The resulting fused feature is then refined by the DPAF module.

Figure 3
Structure of CSGF.

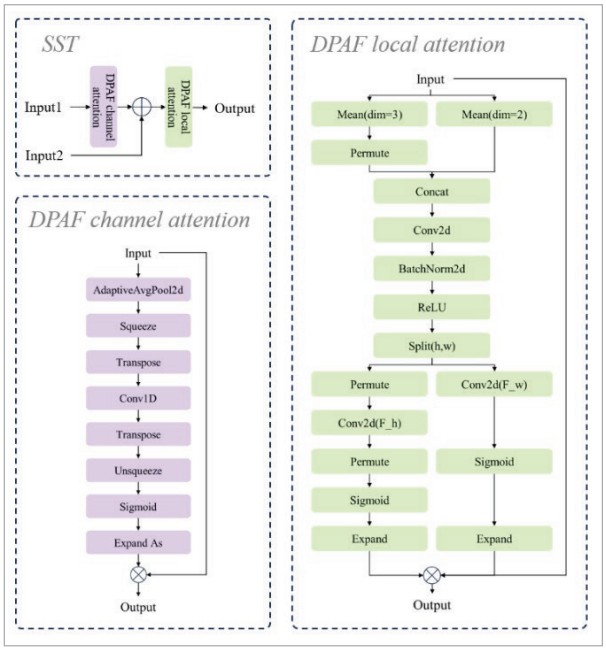


The DPAF module contains two complementary paths: a channel focus path and a spatial focus path. The channel focus path uses global pooling and lightweight channel interaction to estimate channel-wise importance, while the spatial focus path models horizontal and vertical spatial dependencies to highlight target-related regions. The outputs of the two paths are combined to enhance discriminative object responses and suppress irrelevant background interference. This sequential design forms a hierarchical interaction pipeline: higher-level cross-scale fusion, adaptive focus refinement, shallow-detail reinjection, and final multi-scale prediction.

In terms of parameter design in Figure 3, the main learnable components of the hierarchical fusion stage are the channel-alignment layers and the scale-interaction transformations used in DynamicScalSeq. Excluding the internal parameters of the dynamic sampling and alignment operations, the additional parameter overhead is mainly determined by the input channel widths and the unified channel dimension. This design enables explicit cross-scale interaction while keeping the additional learnable structure relatively compact.

Here, the fused feature from the hierarchical fusion stage is further fed into a refinement head for contextual interaction and adaptive focus enhancement. The detailed architecture of this refinement process is illustrated in Figure 4. The contextual interaction stage takes a sequence of aligned multi-scale features as input and models latent depen-

Figure 4
Structure of the fusion-refinement head in DHAF-YOLO, including hierarchical contextual interaction and Dual-Path Adaptive Focus (DPAF).



dencies among them across both scale and spatial dimensions. The output is then aggregated by an additive operation and fed into the next module. The Dual-Path Adaptive Focus module contains a channel focus path and a local focus path. In the channel focus path, AdaptiveAvgPool2d is used to compress the spatial dimensions, and then the Squeeze, Transpose, and Conv1D operations are used to determine the channel relationship. After that, the Transpose, Unsqueeze, and Sigmoid operations are used to obtain the channel-wise weights. Finally, expand and Multi-input operations are used to emphasise the input features. In the local focus path, Mean, Permute, and Concat operations are used to aggregate the spatial information horizontally and vertically. After that, Conv2d, BatchNorm2d, ReLU, and Split operations are used to learn the spatial information, which is then fed into two sub-branches. In the first sub-branch, Conv2d, Permute, and Sigmoid operations are used to obtain the horizontal spatial weights. In the second sub-branch, Conv2d, Permute, and Sigmoid operations are used to obtain the vertical spatial weights. These two kinds of spatial weights are then expanded through the Expand operation and finally used to emphasise the input features. This sequential design forms a clear interaction pipeline: cross-scale fusion → contextual enhancement → adaptive focus refinement.

The detailed fusion and adaptive refinement process is described as follows. For the large-scale feature L, two-stream feature pooling is first performed to preserve both prominent responses and average semantic information. For the small-scale feature S, an upsampling operation is used to match the spatial resolution of the middle-scale feature M. The processed features are then fused with the middle-scale feature to generate a unified representation F. After that, channel attention and spatial attention are estimated from complementary branches and adaptively fused to produce the refined feature.

The following equations formally summarise the entire process:

$$F_l^{processed} = AdaptiveMaxPool2d(L) + AdaptiveAvgPool2d(L) \tag{7}$$

$$F_s^{processed} = Upsample(S, mode=nearest) \tag{8}$$

$$F_{fused} = Concat\,(F_l^{processed}, M, F_s^{processed}) \tag{9}$$

$$\alpha_c = Sigmoid(Expand(Unsqueeze(Transpose(Conv1D(Transpose(Squeeze(AdaptiveAvgPool2d(F)))))))) \tag{10}$$

$$F_{local} = ReLU(BatchNo\ rm2d(Conv2d(Concat(Permute(Mean(F, dim=3)))))) \tag{11}$$

$$a_h, a_w = Sigmoid(Conv2d(Split(Permute(F_{local})))) \tag{12}$$

$$F_{refined} = (F \odot Expand(a_c)) \oplus (F \odot (Expand(a_h) \odot Expand(a_w))), \tag{13}$$

where L denotes the large-scale feature, M denotes the medium-scale feature, and S denotes the small-scale feature. *AdaptiveMaxPool2d* and *AdaptiveAvgPool2d* denote adaptive pooling operations; *Upsample* denotes the upsampling operation; *Concat* denotes concatenation; *Squeeze*, *Unsqueeze*, *Transpose*, *Permute*, and *Expand* are tensor transformation operations; *Conv1D* and *Conv2d* denote

convolution layers; *Mean* denotes averaging; and *Split* denotes the splitting operation. a_c represents channel attention, while a_h and a_w denote the spatial attention weights along the height and width directions. The symbols ⊙ and ⊕ denote multiplication and addition, respectively. In this way, the channel and spatial responses are fused adaptively to obtain the refined feature $F_{refined}$.

4. Experimental Results

4.1. Experimental Setup and Evaluation Methodology

The experimental framework was designed to ensure a fair and reproducible evaluation of the proposed DHAF-YOLO architecture. All models were implemented using the PyTorch-based Ultralytics framework and trained under the same settings on a single NVIDIA GeForce RTX 4060 GPU. The software environment ran under Windows, and the available CUDA version was 12.6. No multi-GPU or distributed training strategy was used.

In all experiments, the input images were resized to 640×640 pixels. Training was performed for 300 epochs with a batch size of 32 and 4 data-loading workers. The optimizer was Stochastic Gradient Descent (SGD). Unless otherwise specified, all optimization settings followed the default training configuration provided by the adopted Ultralytics implementation, and no manual retuning of the optimizer-related hyperparameters was performed in the comparative or ablation experiments. The early stopping patience was set to 0, and all experiments were conducted under the same training schedule for fairness.

Data augmentation also followed the default training pipeline of the adopted Ultralytics implementation. Accordingly, color transformation, mosaic augmentation, and other built-in augmentation operations were applied consistently during training. No separate manual adjustment of augmentation hyperparameters was introduced for the compared models, so that the evaluation focused on architectural differences rather than dataset-specific tuning.

DHAF-YOLO follows the detection objective used in the adopted Ultralytics YOLOv11 framework. Although the feature pyramid is enhanced by the proposed DWFC, CSGF, and DPAF modules, the final prediction layer still uses the standard detection loss formulation implemented in the framework. The overall training loss consists of bounding-box regression loss, classification loss, and distribution focal loss. The bounding-box regression term optimizes the overlap between predicted boxes and ground-truth boxes, the classification term supervises category prediction, and the distribution focal loss refines bounding-box localization by modeling the discrete distribution of box coordinates. The total loss is formulated as:

$$L=\lambda box\ Lbox+\lambda cls\ Lcls+\lambda dfl\ Ldfl, \tag{14}$$

where *Lbox*, *Lcls*, and *Ldfl* denote the bounding-box regression loss, classification loss, and distribution focal loss, respectively. λbox, λcls, and λdfl are the corresponding loss weights, which follow the default settings of the adopted Ultralytics implementation.

For label assignment, the default task-aligned assignment strategy in the Ultralytics YOLOv11 framework was used. During inference, non-maximum suppression was applied to remove redundant detections. Unless otherwise specified, the confidence threshold and IoU threshold for NMS followed the default validation settings of the adopted framework. The same loss formulation, label assignment strategy, and inference settings were used for the baseline, ablation models, and compared YOLO-series detectors to ensure fairness and reproducibility.

Experiments were conducted on the VEDAI-Coco dataset, which is publicly available at Kaggle(https://www.kaggle.com/datasets/kambojharyana/vedai-dataset-coco) and provided in a COCO-style annotation format. To ensure consistency and reproducibility, this study directly followed the released dataset organization and annotation format for the baseline, ablation, and comparative evaluations. Since the present work does not introduce a re-annotation or re-splitting procedure, we avoid restating dataset statistics that were not independently re-verified in this study.

For evaluation, we adopted standard object detection metrics, including Precision, Recall, mAP@50, and mAP@50:95. Here, mAP@50 denotes the mean

Average Precision at an IoU threshold of 0.50, while mAP@50:95 represents the average AP computed over IoU thresholds from 0.50 to 0.95 with a step size of 0.05. These metrics jointly reflect the detection performance of the model under different localization strictness levels.

4.2. Ablation Study Analysis

To verify the effectiveness of the proposed design, we conducted a stage-wise ablation study under a unified 300-epoch training setting. Considering that the cross-scale fusion and adaptive focus modules are structurally coupled in the implemented architecture, the ablation was performed at the level of functional design stages rather than every named submodule in complete isolation. Specifically, starting from the YOLOv11 baseline, we first evaluated the effect of DWFC in the backbone. We then introduced the hierarchical fusion pipeline in the neck/head to assess the contribution of cross-scale interaction and shallow-detail reinjection. After that, DPAF was added on top of the fusion pipeline to evaluate the effect of adaptive focus refinement. Finally, all components were integrated to form the complete DHAF-YOLO framework.

As shown in Table 1, the proposed design stages improve detection performance from different perspectives. Adding DWFC to the baseline mainly enhances local feature extraction and fine-grained response generation, which is beneficial for strengthening weak target representation. Introducing the hierarchical fusion pipeline further improves multi-scale feature interaction by aligning and aggregating pyramid features across different resolutions while preserving high-resolution spatial details. Based on the fusion-enhanced representation, DPAF provides additional refinement by emphasizing target-related channel and spatial responses and suppressing background interference.

When all components are integrated, DHAF-YOLO achieves the strongest overall performance among the tested ablation configurations in terms of Recall, mAP@50, and mAP@50:95. These results indicate that the proposed backbone enhancement, hierarchical fusion, and adaptive focus refinement form an effective feature enhancement pipeline for remote-sensing small-object detection.

4.3. Comparative Performance Evaluation

We further evaluated DHAF-YOLO against several mainstream YOLO-series detectors, ranging from YOLOv5 to YOLOv12, in order to demonstrate the effectiveness of the proposed design. All models were trained using a unified strategy and the same data processing pipeline, so that the comparison focuses on the detector architectures themselves.

The results in Table 2 show that the proposed method achieves the strongest performance in terms of Recall, mAP@50, and mAP@50:95 among the compared detectors. DHAF-YOLO obtains the highest mAP@50:95, surpassing YOLOv9 by 5.0 percentage points. In addition, DHAF-YOLO achieves the highest Recall, indicating that it can recover more small objects missed by competing models. However, DHAF-YOLO does not obtain the highest Precision; YOLOv9 achieves a higher Precision of 85.7%, while DHAF-YOLO obtains 81.7%.

It should be noted that DHAF-YOLO does not dominate all evaluation metrics. Compared with YOLOv9, DHAF-YOLO improves Recall from 93.2% to 95.6%, mAP@50 from 97.7% to 98.3%, and mAP@50:95

Table 1

Ablation Study: Performance Comparison of Different Model Configurations on the VEDAI-Coco Dataset.

Model	Precision(%)	Recall(%)	mAP50(%)	mAP50-95(%)
Baseline	79.7	90.5	95.2	72.4
Baseline + DWFC	89.1	93.0	97.4	73.9
Baseline + Hierarchical Fusion	79.2	91.5	96.7	72.6
Baseline + Fusion + DPAF	81.2	94.8	96.8	76.3
DHAF-YOLO	81.7	95.6	98.3	79.5

Table 2
Benchmarking Detection Performance of YOLO-Series Detectors on the VEDAI-Coco Dataset under a Unified Training and Evaluation Protocol.

Model	Precision(%)	Recall(%)	mAP50(%)	mAP50-95(%)
YOLOv5	77.3	85.9	93.3	72.0
YOLOv6	71.1	83.9	85.9	72.6
YOLOv8	73.9	90.9	93.1	68.9
YOLOv9	85.7	93.2	97.7	74.5
YOLOv10	59.7	88.4	88.7	70.1
YOLOv11	79.7	90.5	95.2	72.4
YOLOv12	64.7	71.5	72.2	55.4
DHAF-YOLO	81.7	95.6	98.3	79.5

from 74.5% to 79.5%, but its Precision decreases from 85.7% to 81.7%. This indicates a recall–precision trade-off. For remote-sensing small-object detection, where missed detections of weak and tiny targets are often critical, the improved Recall and mAP suggest that DHAF-YOLO is more effective at recovering difficult targets. However, in applications where false positives are more costly, the higher Precision of YOLOv9 may still be preferable.

Overall, the results demonstrate that the combination of dynamic local feature extraction, cross-scale gated fusion, and hierarchical attention refinement can effectively improve remote-sensing small object detection. In particular, DHAF-YOLO is more suitable for challenging scenes involving numerous small targets, large scale variation, and complex background interference.

Although DHAF-YOLO is not specifically designed as a lightweight detector, its computational profile still deserves discussion. A full quantitative complexity benchmark, such as Params, GFLOPs, or FPS, is not provided in the current version because this study primarily focuses on detection accuracy and module effectiveness under a unified benchmark setting. To avoid unsupported deployment-oriented claims, we restrict the present discussion to a qualitative characterization of the computational profile. Relative to the baseline architecture, the proposed model introduces additional operations through dynamic convolution, cross-scale interaction, and attention-based refinement, while achieving clear gains in detection accuracy, especially for small-object scenarios. Therefore, the proposed method should be regarded as an accuracy-oriented detector, and its deployment-oriented complexity evaluation will be further examined in future work.

5. Conclusion

In this study, we presented DHAF-YOLO, a remote-sensing small object detector based on dynamic hierarchical aggregation and attention fusion. Specifically, the proposed framework integrates dynamic weighted feature extraction, cross-scale gated fusion, and hierarchical contextual refinement to improve the representation of weak small targets under complex backgrounds.

Extensive experiments and ablation studies show that DHAF-YOLO achieves higher Recall and mAP than the baseline YOLOv11 and the compared YOLO-series detectors, while YOLOv9 still maintains the highest Precision. These results indicate that DHAF-YOLO is particularly effective in reducing missed detections and improving small-object localization accuracy.

Two directions may be explored in future work. First, the current dynamic mechanism can be further extended to architecture-level optimization, such as neural architecture search, to improve adaptability under different remote-sensing scenarios. Second, it is worthwhile to investigate the generalization of

the proposed modules on more compact backbones, as well as their performance in more challenging settings such as extreme small-object detection and video-based remote-sensing detection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Data Sharing Agreement

The dataset used in this study is publicly available at Kaggle: https://www.kaggle.com/datasets/kambojharyana/vedai-dataset-coco.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Chen, H., Wang, Q., Ruan, W., Zhu, J., Lei, L., Wu, X., Hao, G. ALFPN: Adaptive Learning Feature Pyramid Network for Small Object Detection. International Journal of Intelligent Systems, 2023, 1-14. https://doi.org/10.1155/2023/6266209
2. Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., Han, J. Anchor-Free Oriented Proposal Generator for Object Detection. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-11. https://doi.org/10.1109/TGRS.2022.3183022
3. Fang, Q., Lin, Z., Wang, Z. An Efficient Feature Pyramid Network for Object Detection in Remote Sensing Imagery. IEEE Access, 2020, 8, 93058-93068. https://doi.org/10.1109/ACCESS.2020.2993998
4. Fu, S., He, Y., Du, X., Zhu, Y. Anchor-Free Object Detection in Remote Sensing Images Using a Variable Receptive Field Network. EURASIP Journal on Advances in Signal Processing, 2023, 2023, 1-19. https://doi.org/10.1186/s13634-023-01013-2
5. Han, W., Chen, J., Wang, L., Feng, R., Li, F., Wu, L., Tian, T., Yan, J. Methods for Small, Weak Object Detection in Optical High-Resolution Remote Sensing Images: A Survey of Advances and Challenges. IEEE Geoscience and Remote Sensing Magazine, 2021, 9(1), 8-34. https://doi.org/10.1109/MGRS.2020.3041450
6. Huang, W., Li, G., Chen, Q., Ju, M., Qu, J. CF2PN: A Cross-Scale Feature Fusion Pyramid Network Based Remote Sensing Target Detection. Remote Sensing, 2021, 13(5), 847. https://doi.org/10.3390/rs13050847
7. Huangfu, P., Dang, L. A Multi-Scale Pyramid Feature Fusion-Based Object Detection Method for Remote Sensing Images. International Journal of Remote Sensing, 2023, 44, 7790-7807. https://doi.org/10.1080/01431161.2023.2288947
8. Hui, Y., Wang, J., Li, B. DSAA-YOLO: UAV Remote Sensing Small Target Recognition Algorithm for YOLOv7 Based on Dense Residual Super-Resolution and Anchor Frame Adaptive Regression Strategy. Journal of King Saud University - Computer and Information Sciences, 2024, 36(1), 101863. https://doi.org/10.1016/j.jksuci.2023.101863
9. Huo, L., Hou, J., Feng, J., Wang, W., Liu, J. Global and Multi-Scale Aggregate Network for Saliency Object Detection in Optical Remote Sensing Images. Remote Sensing, 2024, 16(4), 624. https://doi.org/10.3390/rs16040624
10. Jin, H., Song, Y., Bai, T., Sun, K., Chen, Y. A Novel Dynamic Context Branch Attention Network for Detecting Small Objects in Remote Sensing Images. Remote Sensing, 2025, 17(14), 2415. https://doi.org/10.3390/rs17142415
11. Kadhim, N., Mourshed, M., Bray, M. Advances in Remote Sensing Applications for Urban Sustainability. Euro-Mediterranean Journal for Environmental Integration, 2016, 1, 7. https://doi.org/10.1007/s41207-016-0007-4
12. Lang, K., Cui, J., Yang, M., Wang, H., Wang, Z., Shen, H. A Convolution with Transformer Attention Module Integrating Local and Global Features for Object Detection in Remote Sensing Based on YOLOv8n. Remote Sensing, 2024, 16(5), 906. https://doi.org/10.3390/rs16050906
13. Li, S., Lin, J., Gang, Y., Pan, X. Small Object Detection in Remote Sensing Images Through Multi-Scale Feature Fusion. The Computer Journal, 2025. https://doi.org/10.1093/comjnl/bxaf040
14. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106
15. Liu, C., Zhang, S., Hu, M., Song, Q. Object Detection in Remote Sensing Images Based on Adaptive Multi-Scale Feature Fusion Method. Remote Sensing, 2024, 16(5), 907. https://doi.org/10.3390/rs16050907

16. Liu, J., Li, S., Zhou, C., Cao, X., Gao, Y., Wang, B. SRAF-Net: A Scene-Relevant Anchor-Free Object Detection Network in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-14. https://doi.org/10.1109/TGRS.2021.3124959

17. Qiu, D., Yang, H., Xu, X., Hu, Y., Fang, Z. Research on Atmospheric Turbulence Observation by LiDAR Based on Imaging Detection Technology. IEEE Instrumentation & Measurement Magazine, 2024, 27(3), 37-42. https://doi.org/10.1109/MIM.2024.10505201

18. Ren, Q., Lu, S., Zhang, J., Hu, R. Salient Object Detection by Fusing Local and Global Contexts. IEEE Transactions on Multimedia, 2021, 23, 1442-1453. https://doi.org/10.1109/TMM.2020.2997178

19. Tang, W., Zhang, Y., Zhang, Q., Wang, W., Zhao, B. An Antibackground Interference Object Detection Method for Complex Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2026, 19, 1288-1304. https://doi.org/10.1109/JSTARS.2025.3566550

20. Tian, B., Chen, H. Remote Sensing Image Target Detection Method Based on Refined Feature Extraction. Applied Sciences, 2023, 13, 8694. https://doi.org/10.3390/app13158694

21. Wan, D., Lu, R., Shen, S., Xu, T., Lang, X., Ren, Z. Mixed Local Channel Attention for Object Detection. Engineering Applications of Artificial Intelligence, 2023, 123, 106442. https://doi.org/10.1016/j.engappai.2023.106442

22. Wang, W., Xu, J., Zhang, R. Optimized Small Object Detection in Low Resolution Infrared Images Using Super Resolution and Attention Based Feature Fusion. PLOS ONE, 2025, 20. https://doi.org/10.1371/journal.pone.0328223

23. Wang, X., Han, C., Huang, L., Nie, T., Liu, X., Liu, H., Li, M. AG-YOLO: Attention-Guided YOLO for Efficient Remote Sensing Oriented Object Detection. Remote Sensing, 2025, 17, 1027. https://doi.org/10.3390/rs17061027

24. Wei, X., Li, Z., Wang, Y. SED-YOLO Based Multi-Scale Attention for Small Object Detection in Remote Sensing. Scientific Reports, 2025, 15. https://doi.org/10.1038/s41598-025-87199-x

25. Wu, Q., Wu, Y., Li, Y., Huang, W. Improved YOLOv5s With Coordinate Attention for Small and Dense Object Detection from Optical Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17, 2543-2556. https://doi.org/10.1109/JSTARS.2023.3341628

26. Xie, S., Zhou, M., Wang, C., Huang, S. CSPPartial-YOLO: A Lightweight YOLO-Based Method for Typical Objects Detection in Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17, 388-399. https://doi.org/10.1109/JSTARS.2023.3329235

27. Yang, L., Gu, Y., Feng, H. Multi-Scale Feature Fusion and Feature Calibration with Edge Information Enhancement for Remote Sensing Object Detection. Scientific Reports, 2025, 15. https://doi.org/10.1038/s41598-025-99835-7

28. Yang, Y., Yang, X., Li, J., Zhang, X., Wang, Y., Liu, L., Wang, M., Zhang, C. Small Object Detection in Remote Sensing Images Based on Redundant Feature Removal and Progressive Regression. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62, 1-14. https://doi.org/10.1109/TGRS.2024.3417960

29. Ye, Y., Ren, X., Zhu, B., Tang, T., Tan, X., Gui, Y., Yao, Q. An Adaptive Attention Fusion Mechanism Convolutional Network for Object Detection in Remote Sensing Images. Remote Sensing, 2022, 14, 516. https://doi.org/10.3390/rs14030516

30. Yuan, Z., Gong, J., Guo, B., Wang, C., Liao, N., Song, J., Wu, Q. Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism. Remote Sensing, 2024, 16, 4265. https://doi.org/10.3390/rs16224265

31. Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62, 1-15. https://doi.org/10.1109/TGRS.2024.3363057

32. Zheng, X., Bi, J., Li, K., Zhang, G., Jiang, P. SMN-YOLO: Lightweight YOLOv8-Based Model for Small Object Detection in Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 2025, 22, 1-5. https://doi.org/10.1109/LGRS.2025.3546034

33. Zhou, S., Zhou, H. Detection Based on Semantics and a Detail Infusion Feature Pyramid Network and a Coordinate Adaptive Spatial Feature Fusion Mechanism Remote Sensing Small Object Detector. Remote Sensing, 2024, 16, 2416. https://doi.org/10.3390/rs16132416