IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 November 2025, accepted 16 December 2025, date of publication 22 December 2025,
date of current version 5 January 2026.

*Digital Object Identifier 10.1109/ACCESS.2025.3646852*

RESEARCH ARTICLE

# SWL-YOLO: A Synergistic Feature Fusion Strategy for Small Object Detection in Remote Sensing Images Based on YOLOv11

**JING ZHANG[1,2], MAS RINA BINTI MUSTAFFA[1], (Senior Member, IEEE), FATIMAH KHALID[1], (Member, IEEE), AND ZAINAL ABDUL KAHAR[1], (Member, IEEE)**

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor 43400, Malaysia
[2]Department of General Education, Anhui Vocational College of City Management, Hefei 230000, China

Corresponding authors: Mas Rina Binti Mustaffa (MasRina@upm.edu.my) and Jing Zhang (gs72462@student.upm.edu.my)

This work was supported in part by Anhui Scientific Research Project under Grant 2024AH050031.

**ABSTRACT** To address the challenges of detail loss, feature extraction difficulties, densely distributed small objects, and insufficient feature information in degraded remote sensing images, we introduce SWL-YOLO, a lightweight model built upon YOLOv11. SWL-YOLO incorporates Spatial Adaptive Feature Module (SAFM), Wavelet Downsampling (WDown), and a Large Selective Kernel (LSK) mechanism to adaptively enhance both spatial and contextual representations. Specifically, the SAFM improves sensitivity to fine-grained spatial features, thereby improving its ability to perceive small targets and edges. The wavelet downsampling module performs wavelet decomposition and subsampling, preserving high-frequency detail information while reducing computational complexity. The LSK mechanism dynamically adjusts receptive fields, enabling the model to better handle small objects, complex backgrounds, and multi-category targets through spatially adaptive feature enhancement and context-aware scale selection. While SAFM ensures enhanced local feature modulation, LSK complements it by providing global context awareness, together forming a synergistic spatial feature fusion mechanism. Furthermore, building upon the CIoU of YOLOv11, we develop an improved GeoCIoU loss, which employs a dual-penalty mechanism for loss calculation to achieve more accurate training feedback. Experiments on the VisDrone and NWPU VHR-10 datasets indicate that SWL-YOLO outperforms the baseline models, with mAP50 improvements of 5.1 % and 4.2 %, respectively, showcasing its superior performance in remote sensing target detection.

**INDEX TERMS** Object detection, remote sensing images, YOLO, spatially-adaptive feature module, large selective kernel mechanism.

## I. INTRODUCTION

Remote Sensing object detection (RSOD) serves as a critical application of remote sensing technology across domains such as geology, forestry, agriculture, and military. However, remote sensing images (RSIs) typically introduce additional challenges, including complex background textures, large variations in object scale, generally low image resolution and densely distributed small targets [1], which significantly hinder the performance of detection algorithms. In recent years, deep learning–based methods have demonstrated clear advantages over conventional approaches, achieving markedly higher detection accuracy.

Among CNN-based RSOD algorithms, existing methods can generally be divided into two major categories: two-stage and single-stage detectors. Two-stage methods, represented by the Fast R-CNN [2], Faster R-CNN [3], and Mask R-CNN [4] series tend to suffer from high model complexity, a large number of parameters, and slow inference speeds. In contrast, single-stage algorithms, such as the YOLO [5], [6], [7], [8], [9] series, RetinaNet [10], and SSD [11], adopt relatively simpler architectures with fewer parameters and

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

faster inference speeds, facilitating their broader application and extensive study within the RSOD domain.

Detecting small objects remains particularly challenging due to limited pixel representation and weak semantic cues. In the context of RSOD, Li et al. [12] enhanced feature fusion in neural networks by incorporating channel and spatial attention mechanisms alongside a weighted bidirectional feature pyramid structure, resulting in an enhanced YOLOv5 framework. To further improve small-object detection, Cui et al. [13] introduced a modified YOLO-based model that leverages low-level features more effectively, although this improvement comes at the expense of reduced performance on large objects. Zhang et al. [14] designed an enhanced YOLOv5 variant incorporating angle regression to more accurately localize arbitrarily oriented objects, achieving tighter bounding boxes but with suboptimal detection accuracy. These methods help mitigate scale imbalance and improve generalization across RSOD. To enable real-time deployment of object detection models in resource-constrained environments, lightweight network designs have gained considerable attention in remote sensing tasks. Zhang et al. [15] developed a lightweight symmetric detection head to replace conventional designs, integrating a C3CA module to strengthen feature extraction and employing FasterConv in the backbone to reduce computational overhead. Addressing multi-scale detection in complex scenes, Yi et al. [16] proposed a multi-scale feature fusion network, LMSFFNet, for object density estimation. Their architecture adopts a MobileViT backbone to achieve a lightweight structure while maintaining competitive accuracy. Liang et al. [17] constructed a joint attention-based multi-scale feature enhancement network on YOLOv5 and incorporated rotation angle recognition to mitigate detection challenges in remote sensing. Wu et al. [18] implemented resolution-adaptive representations to enable flexible scale modeling across various image resolutions. Despite these advances, the substantial scale variation of objects in RSIs makes multi-scale representation indispensable. To address this, Li et al. [19] introduced a feature extraction module with dynamic selection mechanisms to adaptively allocate receptive fields and proposed an adaptive feature-weighted FPN to strengthen Fusion of multi-scale features. Su et al. [20] proposed a strip-wise hierarchical attention module (SHAM) for multi-scale feature fusion and background noise suppression, and employed GhostConv to optimize the feature fusion layer, achieving higher detection accuracy with fewer parameters. Liao et al. [21] employed an efficient attention-based lightweight multi-scale module for effective multi-scale feature extraction and replaced nearest-neighbor upsampling with transposed convolution to minimize feature loss. Zhou et al. [22] enhanced multi-scale feature extraction by integrating improved hybrid convolutions into inverted residual blocks and introduced small-scale and medium-scale feature enhancement modules (SFEM and MFEM) to further refine multi-scale detection. Attention mechanisms have also shown strong capability in highlighting salient features while suppressing irrelevant context. Liao and Zhu et al. [23] strengthened the model's capacity to capture global features and their interdependencies using self-attention. Although this approach demonstrated improvements in detection accuracy, the computational overhead incurred by its sophisticated feature extraction modules limited its practicality. Similarly, Wu et al. [24] utilized deformable convolutional channel attention to prioritize salient features of target objects while suppressing background interference. However, the architectural complexity of this method resulted in slower computational speeds. Li et al. [25] introduced a hybrid attention module to improve the model's response to small targets embedded in complex backgrounds. Although this algorithm achieved higher precision in small object detection, its overall accuracy declined when addressing multi-scale objects commonly present in RSIs. To handle densely distributed objects, Xu [26] proposed a novel pixel interpolation technique that encoded bounding box positional information into corresponding feature points, ensuring alignment between critical pixel features and target objects. While this method enhanced bounding box regression accuracy in dense scenes, it also introduced a substantial increase in model parameters. With respect to arbitrary object orientations, Mei et al. [27] mitigated rotational uncertainty by optimizing a novel objective function that incorporates Fisher discriminant and rotation-invariant regularizations. This improved the model's ability to detect objects with varying orientations but added considerable complexity, leading to unstable training. Yang et al. [28] further introduced a loss function designed to minimize boundary localization errors and mitigate abrupt loss fluctuations near object edges; however, its effectiveness in detecting small targets within remote sensing imagery remained limited. These approaches provide crucial enhancements for handling arbitrarily aligned objects and improve bounding box regression in aerial scenarios.

Despite these advancements, existing algorithms still suffer from structural complexity, imbalanced performance, slow inference speeds and heavy computation in RSOD, especially when dealing with densely distributed small objects and target scale varies greatly. To overcome the limitations in single-stage algorithms, this paper proposes an enhanced SWL-YOLO, which integrates wavelet-based downsampling, spatially-adaptive feature module, and large kernel receptive field selection, together with an improved dual-penalty loss function specifically designed RSOD.

Our main contributions are as follows:

1) The SPPF structure in YOLOv11 has been enhanced with a novel spatial adaptive feature module (SAFM). This module employs spatially-adaptive dynamic weighting to perform region-level and pixel-level feature map modulation, thereby significantly improving the model's ability to capture small objects and edge-level details;

2) The conventional convolution modules in the backbone network have been replaced with a Wavelet Downsampling (WDown) module. This modification is designed to preserve fine-grained spatial details while simultaneously reducing computational overhead;
3) The LSK module dynamically adjusts receptive fields to meet the heterogeneous contextual demands of diverse targets in RSIs. SAFM selectively enhances region-level features, which are further refined through LSK's adaptive multi-kernel aggregation. This cooperative design reduces redundancy while amplifying the discriminative patterns essential for remote sensing detection.
4) Building upon the CIoU [29] loss function in YOLOv11, we propose an advanced Geometric CIoU (GeoCIoU) strategy. This enhanced approach incorporates a dual-penalty mechanism that computes loss based on both bounding box geometry (shape and size) and spatial alignment, providing more accurate training feedback regarding object orientation, scale, and overlap in remote sensing imagery.

## II. PROPOSED METHOD

### A. SWL-YOLO OVERALL STRUCTURE

YOLO is a representative one-stage algorithm for object detection based on CNNs and has long been favored for real-time applications. Despite multiple revisions and updates, it maintains a leading position in the field of object detection. In 2024, Ultralytics released YOLOv11 [30], one of the most advanced versions in the YOLO series, featuring significant innovations in both network architecture and training strategies. The basic framework of YOLOv11 comprises three main components: Backbone, Neck, and Head. Among variants, YOLO11s represents a balanced version in terms of accuracy and architectural design, making it an ideal foundation for reducing parameter count and computational load. The backbone is primarily responsible for extracting the image feature, and includes modules such as Conv, C3K2 [31], SPPF [32] and C2PSA [9]. The neck refines feature maps across multiple layers to generate more discriminative representations, while the head employs depth-wise separable convolutions to reduce redundant connections.

Despite introducing additional modules, the overall computational cost remains low due to efficient architectural design. Specifically, the SAFM utilizes channel splitting, multi-branch pooling, and $1 \times 1$ convolutions to perform spatial modulation without introducing excessive learnable parameters. Additionally, parallel pooling and attention-generation layers in LSK are shared across spatial channels, further minimizing redundancy. Additionally, parallel pooling and attention generation layers in LSK are shared across spatial channels, further minimizing redundancy. As demonstrated in the ablation study (Table 1), integrating SAFM and LSK only slightly increases computation from 21.3 to 25.7 GFLOPs while reducing parameter size from 9.41M to 8.45M, highlighting that the proposed enhancements maintain computational efficiency and support real-time deployment.

This paper addresses several limitations of YOLOv11, including challenges in detecting small objects, missed and false detections, and loss of detail in low-quality images. Based on YOLO11s, we propose SWL-YOLO for RSIs, whose overall architecture is illustrated in FIGURE 1.

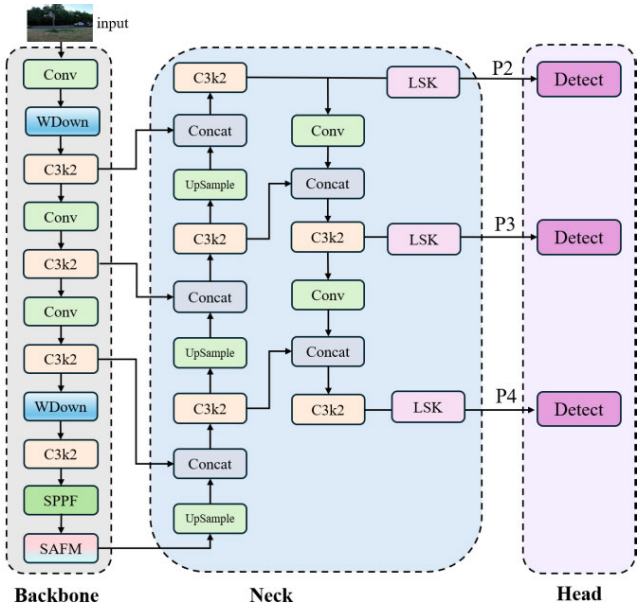



**FIGURE 1.** The overview of the proposed SWL-YOLO model.

### B. SAFM-BASED FEATURE FUSION MECHANISM

In object detection tasks, multi-scale feature fusion [33] plays a vital role in improving model performance. In YOLO11s, the backbone leverages an SPPF combined with the C2PSA architecture to integrate high-level features rich in semantic information with low-level features containing detailed spatial cues. However, this design introduces substantial computational overhead while offering limited improvement for small-object detection. To address this, the spatial adaptive feature module (SAFM) module employs multi-scale pooling operations to aggregate features with diverse receptive fields, thereby enhancing contextual information representation. In this work, we propose an enhanced SPPF structure for YOLO11s that incorporates SAFM after the multi-scale pooling operations. SAFM dynamically adjusts feature map weights and structures according to spatial locations in the input image, using spatial-aware dynamic weighting to refine both region-level and pixel-level features. This design significantly improves the model's ability to detect small objects and capture edge details, while also increasing robustness in challenging scenarios, such as occlusion or objects of varying scales.

As illustrated in FIGURE 2, the SAFM employs a feature pyramid to generate spatially adaptive module parameters for

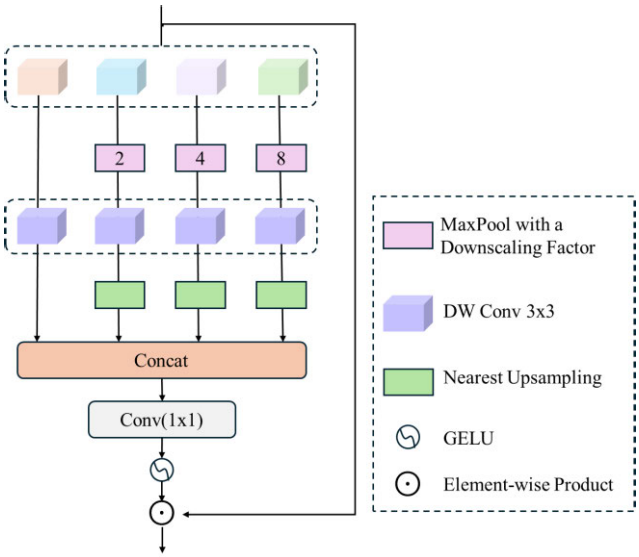



**FIGURE 2.** Spatially-adaptive feature module.

input feature maps. To reduce computational overhead, the module first partitions the channels into four components for separate processing of local and global information using $\boldsymbol{Split}(\cdot)$ operation. For the first partitioned component, a depthwise convolution layer utilizing a 3 × 3 kernel is directly applied. The remaining three components undergo downsampling $\boldsymbol{DW-Conv}_{3\times3}(\cdot)$ through $3\times3$ convolutional layers with strides of 2, 4, and 8 respectively to minimize computational costs, followed by upsampling to the original resolution via nearest interpolation.

$$\begin{aligned}[\boldsymbol{X}_0, \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3] &= \boldsymbol{Split}(\boldsymbol{X}),\\ \hat{\boldsymbol{X}}_0 &= \boldsymbol{DW}-\boldsymbol{Conv}_{3\times3}(\boldsymbol{X}_0),\\ \hat{\boldsymbol{X}}_i &= \uparrow_p(\boldsymbol{DW}-\boldsymbol{Conv}_{3\times3}(\downarrow_{\frac{p}{2^i}}(\boldsymbol{X}_i))),\ 1\le \boldsymbol{i}\le 3,\end{aligned} \tag{1}$$

Subsequently, the processed features are concatenated and integrated via a 1 × 1 convolutional layer, enabling synergistic utilization of both local and global features. It can be expressed by:

$$\hat{\boldsymbol{X}} = \boldsymbol{Conv}_{1\times1}(\boldsymbol{Concat}([\hat{\boldsymbol{X}}_0, \hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2, \hat{\boldsymbol{X}}_3])) \tag{2}$$

Finally, the reconstructed features undergo normalization through the GELU [34] activation function $\boldsymbol{\phi}\left(\hat{\boldsymbol{X}}\right)$, followed by element-wise multiplication with the modulation parameters. It can be written as:

$$\bar{\boldsymbol{X}} = \boldsymbol{\phi}(\hat{\boldsymbol{X}}) \odot \boldsymbol{X} \tag{3}$$

In this process, feature diversity is enhanced through dimensionality reduction using pooling operations, channel partitioning, and diverse convolutional operations, followed by feature integration via concatenation to strengthen the final representation. The proposed architecture effectively addresses key challenges in RSOD, particularly the irreversible loss of fine-grained spatial details that occurs during conventional downsampling, which can severely impair the detection of geometrically sensitive features such as edges, corners, and textural patterns. Additionally, it alleviates the limited discriminative capability for multi-scale targets, a common issue in optical remote sensing imagery, thereby improving detection performance across objects of varying sizes and spatial complexities.

### C. WAVELET DOWNSAMPLING CONVOLUTION

Remote sensing imagery is inherently limited by technical costs, atmospheric conditions, and acquisition altitudes, often resulting in degraded image quality characterized by low resolution, complex backgrounds, and particularly small targets with limited pixel coverage. These small targets primarily convey their features through edge and texture details rather than holistic shapes. Existing small target detection methods continue to struggle with challenges such as insufficient contrast and significant loss of fine-grained details. In this context, the discrete wavelet transforms (DWT) [35] provide a theoretically grounded multi-scale analytical framework, capable of hierarchical feature representation from coarse to fine granularity. DWT is particularly effective for detecting variably sized small targets, as its multi-level decomposition preserves and enhances discriminative characteristics. Unlike conventional downsampling methods such as max pooling operations, DWT maintains critical high-frequency components during decomposition without blurring or discarding essential spatial details. In this work, we implement DWT and inverse discrete wavelet transform (IDWT) as integrated layers specifically tailored for small-target detection. The following section establishes the theoretical foundation of discrete wavelet analysis, beginning with fundamental formulations for one-dimensional and two-dimensional data, with the methodology readily extensible to higher-dimensional cases through analogous mathematical operations.

#### 1) 1D DWT/IDWT

Discrete wavelet transform decomposes 1D data $\boldsymbol{x} = \{\boldsymbol{x}_j\}_{j\in Z}$ into low-frequency subbands $\boldsymbol{x}_{low} = \left\{\boldsymbol{x}_k^{(low)}\right\}_{k\in Z}$ and high-frequency subbands $\boldsymbol{x}_{high} = \left\{\boldsymbol{x}_k^{(high)}\right\}_{k\in Z}$,

$$\begin{cases}\boldsymbol{x}_k^{(low)} = \sum_j \boldsymbol{l}_{j-2k}\boldsymbol{x}_j,\\ \boldsymbol{x}_k^{(high)} = \sum_j \boldsymbol{h}_{j-2k}\boldsymbol{x}_j,\end{cases} \tag{4}$$

In Equation (4), $\boldsymbol{l} = \{\boldsymbol{l}_k\}_{k\in Z}, \boldsymbol{h} = \{\boldsymbol{h}_k\}_{k\in Z}$ represents discrete wavelet of low-pass and high-pass filters. Inverse discrete wavelet transform rebuilds $\boldsymbol{x}$ using $\boldsymbol{x}_{low}, \boldsymbol{x}_{high}$, where

$$\boldsymbol{x}_j = \sum_k \left(\boldsymbol{l}_{j-2k}\boldsymbol{x}_k^{(low)} + \boldsymbol{h}_{j-2k}\boldsymbol{x}_k^{(high)}\right) \tag{5}$$

Equation (4) and Equation (5) can be rewritten as

$$\begin{aligned}\boldsymbol{x}_{low} &= \boldsymbol{L}\boldsymbol{x},\\ \boldsymbol{x}_{high} &= \boldsymbol{H}\boldsymbol{x},\end{aligned}$$

$$x = L^T x_{low} + H^T x_{high}, \tag{6}$$

where

$$L = \begin{pmatrix} \cdots & \cdots & \cdots & & & & \\ \cdots & l_{-1} & l_0 & l_1 & \cdots & & \\ & & \cdots & l_{-1} & l_0 & l_1 & \cdots \\ & & & & & \cdots & \cdots \end{pmatrix}, \tag{7}$$

$$H = \begin{pmatrix} \cdots & \cdots & \cdots & & & & \\ \cdots & h_{-1} & h_0 & h_1 & \cdots & & \\ & & \cdots & h_{-1} & h_0 & h_1 & \cdots \\ & & & & & \cdots & \cdots \end{pmatrix}, \tag{8}$$

Equations (6) denote the forward propagation of DWT and IDWT. Equations (7) and (8) represent the low-pass and high-pass filter matrices respectively. The gradients information for their back propagation can be separately deduced from Equation (5) and Equation (6).

#### 2) WAVELET DOWNSAMPLING MODULE

The proposed wavelet-based downsampling pooling framework follows the computational procedure outlined below: First, the input image undergoes multiscale decomposition via DWT, generating four distinct subbands. A specialized denoising operation is then applied exclusively to the high-frequency detail subbands, aiming to preserve critical edge information while effectively suppressing noise components. After this selective denoising, all subbands are reconstructed using IDWT to maintain feature the integrity of the features. Finally, adaptive threshold segmentation is applied to the reconstructed image, enabling precise detection of small targets within the enhanced feature space. As illustrated in FIGURE 3, the Wavelet downsampling algorithm is as follows:

**Step 1**: Shallow Feature Extraction. The input noisy data $\boldsymbol{X}$ is first processed through shallow feature extraction. A convolutional layer is applied to extract preliminary features, which are then normalized using Batch Normalization and activated via SiLU functions [36]. This produces processed intermediate features $\boldsymbol{F}$ that retain essential spatial information while ensuring a normalized data distribution.

**Step 2**: Multiscale Decomposition. The feature maps $\boldsymbol{F}$ are decomposed via DWT into four distinct subbands: one low-frequency subband $\boldsymbol{X_{ll}}$ capturing global structural information, and three high-frequency subbands $\boldsymbol{X_{lh}}, \boldsymbol{X_{hl}}, \boldsymbol{X_{hh}}$ containing edge and texture components.

In equations (9), the two-dimensional DWT is typically implemented by sequentially applying 1D DWT to each row and column of 2D data $\boldsymbol{X}$, i.e.,

$$\begin{aligned} \boldsymbol{X_{ll}} &= \boldsymbol{L}\boldsymbol{X}\boldsymbol{L}^T, \\ \boldsymbol{X_{lh}} &= \boldsymbol{H}\boldsymbol{X}\boldsymbol{L}^T, \\ \boldsymbol{X_{hl}} &= \boldsymbol{L}\boldsymbol{X}\boldsymbol{H}^T, \\ \boldsymbol{X_{hh}} &= \boldsymbol{H}\boldsymbol{X}\boldsymbol{H}^T \end{aligned} \tag{9}$$

where $\boldsymbol{X_{ll}}$ is the low-frequency portion, which contains the fundamental structure of the object. $\boldsymbol{X_{lh}}, \boldsymbol{X_{hl}}, \boldsymbol{X_{hh}}$ are the three high-frequency part. They represent the horizontal, vertical and diagonal characteristics of the image including most of the noise, respectively.

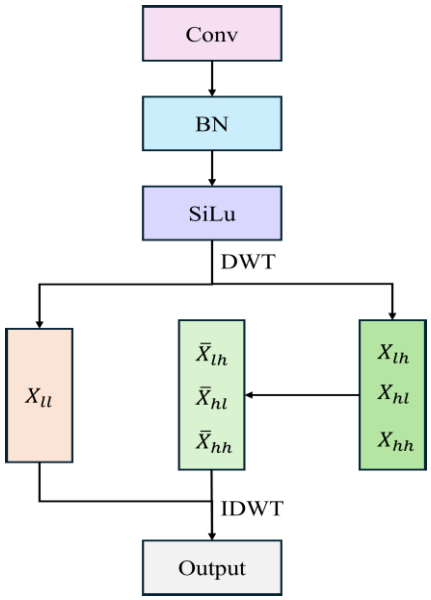



**FIGURE 3.** The wavelet downsampling module.

**Step 3**: High-Frequency Denoising. An adaptive denoising operation is selectively applied to the high-frequency subbands to suppress noise while preserving geometrically significant features. This process maintains the critical edge information in horizontal, vertical and diagonal orientations $\bar{\boldsymbol{X}}_{\boldsymbol{lh}}, \bar{\boldsymbol{X}}_{\boldsymbol{hl}}, \bar{\boldsymbol{X}}_{\boldsymbol{hh}}$.

**Step 4:** Multiscale Reconstruction. The low-frequency component and the processed high-frequency components are jointly reconstructed through the IDWT, yielding enhanced feature maps that effectively integrate multi-scale frequency-domain information.

In equation (10), the 2D IDWT is executed with

$$\boldsymbol{X} = \boldsymbol{L}^T\boldsymbol{X}_{\boldsymbol{ll}}\boldsymbol{L} + \boldsymbol{H}^T\boldsymbol{X}_{\boldsymbol{lh}}\boldsymbol{L} + \boldsymbol{L}^T\boldsymbol{X}_{\boldsymbol{hl}}\boldsymbol{H} + \boldsymbol{H}^T\boldsymbol{X}_{\boldsymbol{hh}}\boldsymbol{H} \tag{10}$$

### D. LARGE SELECTIVE KERNEL MECHANISM

The theoretical motivation behind our architecture arises from the need to jointly model spatial precision and contextual diversity, two properties that are especially critical for small-object detection. Existing attention and kernel-adaptation mechanisms are often insufficient when applied independently. In remote sensing scenarios, objects are typically extremely small, frequently occluded, embedded in backgrounds with structurally similar interference, and span diverse categories requiring widely varying contextual information. Insufficient long-range contextual information can therefore lead to missed or incorrect detections of minute targets. The LSK mechanism [37] addresses these challenges through a dynamic receptive field approach. It employs large-kernel decomposed convolutions to generate multi-scale features, coupled with a spatially selective mechanism that

adaptively prioritizes the most relevant contextual regions based on input content. By substantially expanding the receptive field, this approach enables the model to capture global environmental patterns with targets, thereby enhancing robustness against complex background interference and scale variations. Moreover, the content-aware spatial selection mechanism significantly improves the model's ability to distinguish genuine targets from structurally similar false positives, particularly in occlusion scenarios.

As illustrated in FIGURE 4, the mechanism first employs large kernel decomposition convolutions (implemented via cascaded depthwise separable convolutions) to generate feature maps with varying receptive fields. LSK employs parallel depthwise convolutions with kernel sizes of $5\times5$, and $7\times7$, chosen to balance computational overhead and context capture across varying object sizes. The decomposed features undergo channel-wise mixing through $1\times1$ convolutions $\boldsymbol{F_i}^{1\times1}$ to fuse cross-channel information. Subsequently, these multi-scale features $\tilde{\boldsymbol{U}}=\left[\widetilde{\boldsymbol{U}_1};\widetilde{\boldsymbol{U}_2}\right]$ are concatenated on the channel dimensions, followed by parallel average pooling and max pooling operations to generate spatial attention descriptors $\boldsymbol{SA}$.

$$\begin{aligned}\boldsymbol{SA_{avg}} &= \boldsymbol{P_{avg}}\left(\widetilde{\boldsymbol{U}}\right)\\ \boldsymbol{SA}_{\max} &= \boldsymbol{P}_{\max}\left(\widetilde{\boldsymbol{U}}\right)\\ \boldsymbol{SA} &= [\boldsymbol{SA_{avg}}, \boldsymbol{SA}_{\max}]\end{aligned} \tag{11}$$

The pooled results are concatenated and passed through a convolution layer $\boldsymbol{F}^{2\to N}$ that outputs $\boldsymbol{N}$ spatial attention maps $\boldsymbol{SA}$.

$$\boldsymbol{SA} = \boldsymbol{F}^{2\to N}\left(\left[\boldsymbol{SA_{avg}};\boldsymbol{SA}_{\max}\right]\right) \tag{12}$$

Sigmoid activation $\boldsymbol{SA}_i$ is applied to each spatial attention map $\boldsymbol{SA}$ to generate an independent spatial selection mask for every decomposed large kernel. The context augmented representation is obtained by combining the multi-scale feature with the corresponding spatial attention weight to get the weighted sum. Channel integration and dimensionality reduction are subsequently achieved through convolutional fusion $\boldsymbol{F}$, yielding compact contextual features $\boldsymbol{S}$.

$$\boldsymbol{S} = \boldsymbol{F}(\sum_{i=1}^{N}\boldsymbol{SA_i}\cdot\boldsymbol{U_i}) \tag{13}$$

The fused features are subsequently element-wise multiplied with the original input $\boldsymbol{X}$ to generate the final output feature map. This LSK framework effectively achieves efficient multi-scale context modeling and spatially adaptive feature enhancement, substantially improving both accuracy and robustness in RSOD by adaptively emphasizing critical spatial regions while suppressing redundant or interfering patterns.

Given the diverse spatial distributions and scales of targets in RSIs, the LSK module's ability to dynamically highlight large-scale spatial context without sacrificing fine-grained detail is crucial for accurately detecting small objects in cluttered scenes. SAFM provides detailed local spatial features, while LSK dynamically adjusts the context scope. Operating on the refined features produced by SAFM, LSK dynamically modulates the receptive field, forming a coherent mechanism for joint local and global feature enhancement.

### E. IMPROVED LOSS FUNCTION

Regression serves as a fundamental task, where the regression loss function significantly influences model performance. In the YOLO11s model, CIoU loss function serves as the fundamental regression loss, calculated in equation (14). The computation process is as follows:

$$\boldsymbol{CIoU} = \boldsymbol{IOU} - \frac{\boldsymbol{d}^2}{\boldsymbol{c}^2} - \boldsymbol{\alpha v} \tag{14}$$

Here, $d$ denotes the distance between the centroids of the predicted and ground-truth bounding boxes, while $c$ is the diagonal length of their minimum enclosed rectangle. $\alpha$ and $v$ are parameters governing the aspect ratio. This loss function comprehensively optimizes detection performance by incorporating multiple geometric factors, including overlap area, aspect ratio discrepancy and centroid distance.

However, when applied to remote sensing imagery with substantial shape variations, CIoU exhibits limitations due to its inability to capture directional information and its suboptimal handling of overlap. To resolve these concerns, we propose GeoCIoU, a novel IoU-based metric that extends CIoU by incorporating an additional penalty term. Specifically, GeoCIoU reduces the distances between predicted and actual bounding boxes. This dual-penalty mechanism, which combines the original CIoU components with the new corner-distance constraints, provides more accurate training feedback. The complete computational procedure is detailed in Equation (15):

$$\begin{aligned}\boldsymbol{GeoCIoU} = \boldsymbol{IoU} - \boldsymbol{\omega}_1 \times \left(\frac{\boldsymbol{d}^2}{\boldsymbol{c}^2} + \boldsymbol{\alpha v}\right) - \boldsymbol{\omega}_2\\ \times \left(\frac{\boldsymbol{d}_1^2}{\boldsymbol{w}_{gt}^2 + \boldsymbol{h}_{gt}^2} + \frac{\boldsymbol{d}_2^2}{\boldsymbol{w}^2 + \boldsymbol{h}^2}\right)\end{aligned} \tag{15}$$

Figure 5 illustrates the schematic diagram of GeoCIoU, where the green and red bounding boxes stand for actual and predicted boxes, respectively. $d_1$ and $d_2$ represent the distances from the bottom-left corners to top-right corners of the predicted and actual bounding boxes, respectively. The coefficients $\omega_1$ and $\omega_2$ denote the weighting factors for these two penalty terms, whose optimal values will be empirically determined and discussed in the experimental section.

## III. EXPERIMENTS

### A. DATASETS

To evaluate the detection performance, our proposed model was primarily tested on the Visdrone2019 [38] and NWPU VHR-10 [39] datasets.

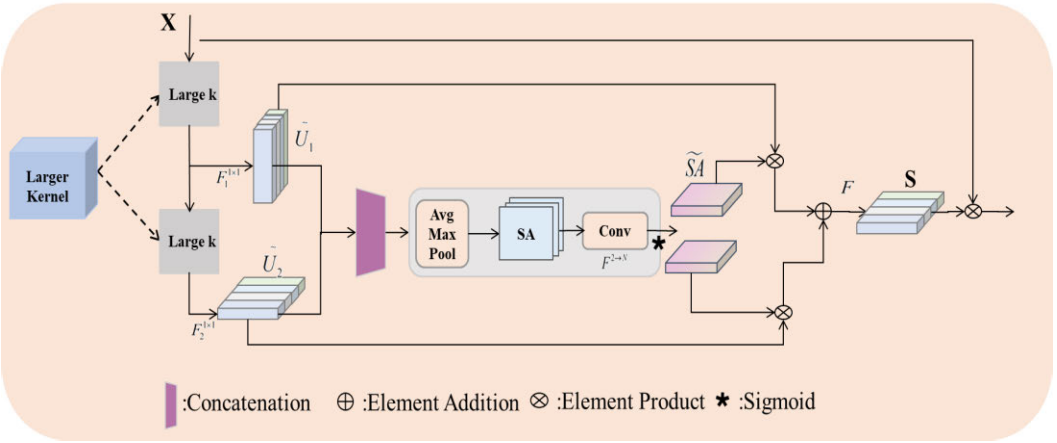



**FIGURE 4.** Large selective kernel mechanism.

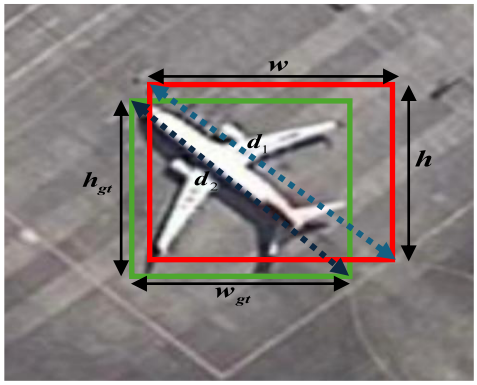



**FIGURE 5.** Schematic diagram of GeoCIoU.

The VisDrone2019 dataset was collected by drones equipped with cameras at various locations and altitudes, comprising 10,209 static images. It includes 10 target categories: pedestrian, individual, bicycle, automobile, van, tricycle, truck, awning-tricycle, bus, and motorcycle.

The NWPU VHR-10 dataset contains 10 categories for geospatial remote sensing object detection, consisting of 800 images in total—650 with targets and 150 background images. Categories include airplanes, ships, storage container, baseball diamond, tennis court, basketball court, running track, seaport, bridge, and motor vehicle.

While NWPU VHR-10 offers rich target diversity, its relatively small size (800 images) can lead to overfitting without extensive data augmentation. Additionally, object sizes are highly imbalanced, with some categories underrepresented. As illustrated in FIGURE 6, to mitigate poor generalization resulting from the limited size of the NWPU VHR-10 dataset, we applied a series of data augmentation techniques, including flipping, rotation, noise addition, fog simulation, and HSV conversion. To ensure that the augmented dataset preserved the statistical properties of the original data, we conducted a distribution analysis comparing key characteristics between the original and augmented datasets, confirming that all categories maintained their relative frequencies within ±2% variation. The original dataset was first split into training, validation, and testing sets in a 7:2:1 ratio. Data augmentation operation was then applied exclusively to the training set, resulting in an expanded dataset of over 3,000 images. The validation and testing sets remained unchanged to ensure fair and unbiased performance evaluation. This partitioning strategy preserves the original class balance while providing sufficient samples to support robust model training and reliable evaluation.

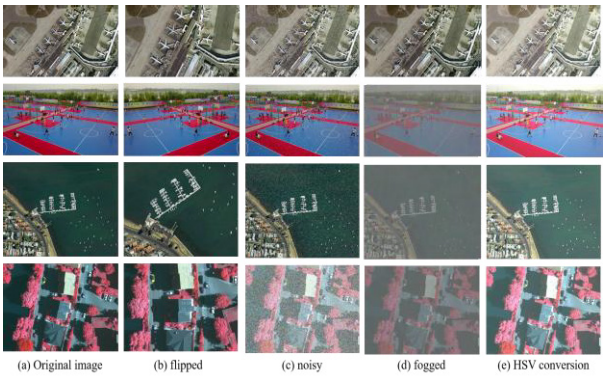



**FIGURE 6.** Data expansion operations.

Figure 7 illustrates the joint distribution of target center locations (normalized $x$, $y$ coordinates) and scales (width and height) in the augmented dataset. The $x$ and $y$ axes represent normalized spatial positions (ranging from 0.0 to 1.0), while the color gradient or point density indicates target frequency. The subplots along the diagonal show the individual distributions of width and height (ranging from 0.0 to 0.5), revealing a concentration of smaller-scale targets (width <0.3, height <0.2).

To further assess model performance across different object categories, the normalized confusion matrix is presented in FIGURE 8. This matrix offers an in-depth

perspective on classification accuracy and inter-class confusion. The model demonstrates strong discriminative capability for several categories, particularly for car and background, achieving correct classification rates of 0.79 and 0.84, respectively. These results indicate that the model effectively captures distinctive features for these classes.

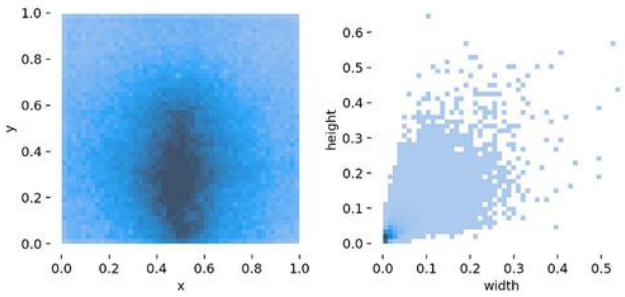



**FIGURE 7.** Distribution of target center locations and scales in the enhanced dataset.

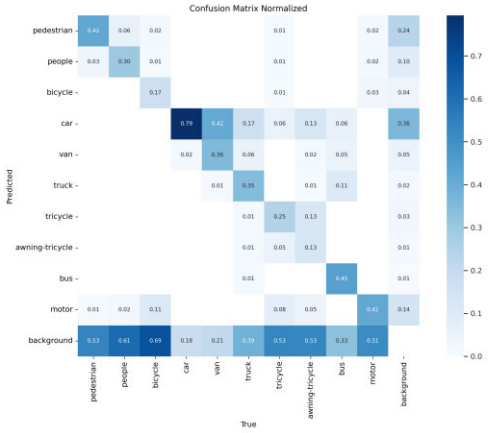



**FIGURE 8.** Confusion matrix of the proposed model in NWPU VHR-10 dataset.

### B. IMPLEMENTATION DETAILS

To ensure fair and reproducible comparisons, all baseline and proposed models were trained and evaluated under the same hardware environment. Experiments were conducted on an Ubuntu 20.04 platform, with an AMD EPYC 7T83 64-Core CPU and an NVIDIA GeForce RTX 3080 GPU. Model development and code execution were performed using the PyCharm integrated development environment on Windows 10. The implementation was based on PyTorch 1.10.0, with CUDA 11.3 enabling GPU parallel processing and cuDNN 8.3.4 serving as the acceleration library to enhance training efficiency. The Python version used was 3.8. The experimental parameters were set as follows: a batch size of 16, training for 100 epochs, and an initial learning rate of 0.01. Additional optimization parameters included a momentum coefficient of 0.937 and a weight decay rate of 0.0005 during model initialization. Hyperparameter initialization followed the prior YOLO11s configuration [30], which is optimized for real-time applications, while final values were empirically tuned on the VisDrone validation set to balance convergence stability and accuracy.

### C. EVALUATION METRICS

The assessment metrics for object detection algorithms can be broadly divided into model complexity and detection accuracy. Model complexity is typically quantified by computational load (GFLOPs) and parameter count (Params), with higher values indicating increased complexity. Detection performance is commonly evaluated using Precision (P), Recall (R), mean Average Precision (mAP), GFLOPs, and parameter count, providing a comprehensive assessment of the model. The above indicators provide an overall evaluation for the model. Precision, as defined in Equation (16), measures the proportion of correctly identified positive instances among all predicted positives. Recall, given in Equation (17), represents the ratio of correctly identified positive instances to all actual positive instances. The parameter count corresponds to the total number of learnable parameters across all network layers during training. The confusion matrix serves as the foundation for computing Precision, Recall, and mAP, enabling precise evaluation of detection performance. The precision and recall are detailed as follows:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{16}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{17}$$

True Positives (TP) is the number of positive samples correctly classified by the model. False Positives (FP): Number of false positives are negative sample which labeled as positive. False Negatives (FN) is where positives were missed by the model. In (18), individual class AP is formally defined as the area under the Precision-Recall (PR) curve, which is mathematically defined as:

$$\boldsymbol{AP} = \int_0^1 \boldsymbol{P(R)dR} \tag{18}$$

mAP represents the average of the AP values, reflecting a summary of the model's complete detection level, calculated in Equation (19). $mAP_{0.5}$ is computed with a constant IoU threshold of 0.5. $mAP_{0.5:0.95}$ is calculated by using a series of IoU scores over a set of IoU scores ranging from 0.5 to 0.95

$$\boldsymbol{mAP} = \frac{\sum_{i=1}^{N} \boldsymbol{AP}_i}{\boldsymbol{N}} \tag{19}$$

In Equation (19), $N$ denotes the number of categories.

### D. ABLATION EXPERIMENT

To verify the effectiveness of the proposed improvement methods, ablation experiments were conducted on two benchmark datasets: VisDrone and NWPU VHR-10. The

baseline model was YOLO11s. To demonstrate the generalization capability, we also selected a non-remote sensing dataset: CityPersons. The detailed experimental results are summarized in TABLE 1.

**TABLE 1. Ablation experiment on different datasets.**

| Dataset | Method | $mAP_{0.5}$ | $mAP_{0.5:0.95}$ | Acc | #Param | Recall | GFLOPs |
|---|---|---|---|---|---|---|---|
| VisDrone | YOLO11s (A) | 59.9 | 54.0 | 70.5 | 9.41 | 89.2 | 21.3 |
| | A+SAFM (B) | 60.6 | 54.3 | 73.6 | 9.45 | 89.1 | 27.0 |
| | B+WDown (C) | 61.3 | 54.9 | 78.4 | 8.69 | 89.5 | 26.5 |
| | C+LSK(D) | 64.5 | 55.2 | 79.8 | 8.45 | 89.6 | 25.7 |
| | D+GeoCIoU | **65** | 57.1 | 80.3 | 8.36 | 89.8 | 25.4 |
| NWPU VHR-10 | YOLO11s (A) | 82.7 | 81.4 | 75.4 | 11.1 | 86.3 | 28.4 |
| | A+SAFM (B) | 83.2 | 81.9 | 78.5 | 11.6 | 87.6 | 29.5 |
| | B+WDown (C) | 84.5 | 82.5 | 78.9 | 10.4 | 87.9 | 27.8 |
| | C+LSK(D) | 85.3 | 83.4 | 80.6 | 10.3 | 88.1 | 27.7 |
| | D+GeoCIoU | **86.9** | 85.6 | 82.2 | 10.1 | 88.4 | 26.2 |
| CityPersons | YOLO11s(A) | 71.2 | 60.5 | 73.2 | 16.7 | 75.5 | 26.4 |
| | A+SAFM (B) | 71.7 | 60.8 | 73.4 | 16.6 | 76.5 | 27.6 |
| | B+WDown (C) | 72.5 | 61.3 | 73.8 | 16.4 | 76.3 | 28.4 |
| | C+LSK(D) | 73.8 | 62.1 | 74.3 | 16.3 | 76.8 | 28.7 |
| | D+GeoCIoU | **74.9** | 63.4 | 74.5 | 16.1 | 77.1 | 29.5 |

As shown in TABLE 1, the integration of the SAFM, WDown, and GeoCIoU modules leads to improvements across multiple evaluation metrics. The incorporation of the SAFM enhances detection performance, particularly in terms of accuracy, indicating that the SAFM module optimizes feature extraction capability and enables the model to focus more effectively on target boundary details. The introduction of the WDown module significantly boosts both recall and mAP. The increase in recall demonstrates that the model is capable of detecting more positive samples, thereby minimizing missed detections and capturing more relevant information in complex scenarios. The LSK module markedly improves mAP for RSOD while simultaneously reducing computational cost. This improvement in mAP reflects enhanced boundary localization accuracy, effectively addressing the limitations of traditional sampling methods in adapting to target variations. The modified loss function, GeoCIoU, further refines target matching, improving the precision of detection boundaries. After incorporating all enhancements, the proposed model achieves $mAP_{0.5}$ of 65%, 86.9%, and 74.9% on the VisDrone, NWPU VHR-10, and CityPersons datasets, representing increases of 5.1%, 4.2%, and 3.7%, respectively, compared to the baseline. Overall, the model demonstrates significant performance gains while maintaining a reasonable computational cost without introducing excessive overhead.

The ablation experiments conducted on three datasets demonstrate the effectiveness of SWL-YOLO in balancing computational efficiency and detection performance. On the VisDrone dataset, the baseline YOLO11s model has 9.41M parameters and a computational cost of 21.3 GFLOPs. With progressive enhancements, SWL-YOLO reduces the parameter count to 8.36M while slightly increasing GFLOPs to 25.4. On the NWPU VHR-10 dataset, SWL-YOLO achieves even more notable efficiency gains: the parameter count decreases from 11.1M to 10.1M, GFLOPs drop from 28.4 to 26.2, and $mAP_{0.5}$ improves by 4.2%. On the CityPersons dataset, integrating these modules results in significant improvements in $mAP_{0.5}$ (from 71.2% to 74.9%), $mAP_{0.5:0.95}$ (from 60.5% to 63.4%), and Recall (from 76.4% to 78.2%), while maintaining reasonable computational cost, with GFLOPs increasing modestly from 26.4 to 29.5. The reduction in parameters is mainly attributed to the lightweight design of the WDown module and the efficient depthwise separable convolutions in LSK, whereas the slight increase in GFLOPs stems from the multi-branch structure of SAFM. This dual optimization of lower computational cost and higher accuracy highlights the model's scalability and suitability for complex remote sensing scenarios. The design effectively balances feature enhancement with computational efficiency, preserving inference speed suitable for edge deployment.

As shown in FIGURE 9, the enhanced model outperforms the baseline across all key performance metrics, demonstrating superior detection capability and stability. Particularly under challenging conditions such as uneven lighting, severe target occlusion, and coexistence of multi-scale targets, the improved model exhibits strong robustness and adaptability. These results confirm the effectiveness of the proposed modules in complex environments.

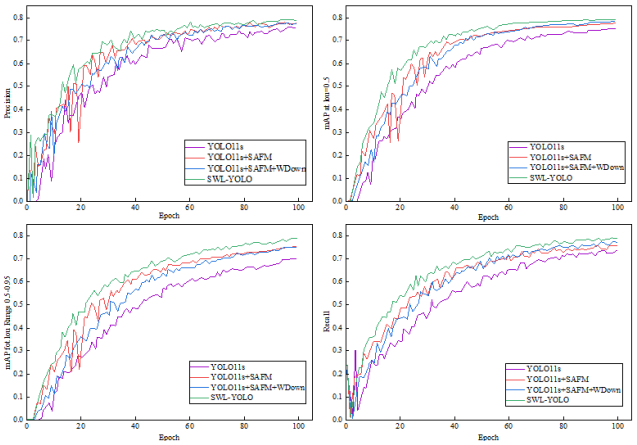



**FIGURE 9. Comparative performance in occluded, low-light, and blurred scenarios (VisDrone subset, metrics: Precision, Recall, $mAP_{0.5}$, $mAP_{0.5:0.95}$).**

The GeoCIoU loss function evaluates the alignment between predicted and actual bounding boxes using two distinct penalty terms. To better balance shape and positional discrepancies, we performed comprehensive tests to investigate the effects of different weight coefficients, $\omega_1$ and $\omega_2$ on detection performance. As demonstrated in TABLE 2, GeoCIoU consistently outperforms CIoU by leveraging its dual-penalty mechanism.

The proposed GeoCIoU loss function employs a dual-penalty mechanism with carefully optimized weight coefficients ($\omega_1 = 0.6$ for geometric shape preservation, $\omega_2 = 0.4$ for spatial alignment). Experimental results reveal

that on the VisDrone dataset (characterized by predominantly small targets<50 pixels), varying $\omega_1$ within the range [0.5,0.7] yields marginal performance fluctuations (<0.8 % mAP variation), indicating remarkable stability for small object detection. Conversely, evaluation shows consistent performance maintenance when $\omega_2$ remains within [0.4,0.6]. This stability arises from the complementary roles of the two penalty terms: $\omega_1$ primarily enhances edge localization accuracy for small objects by preserving high-frequency features, while $\omega_2$ maintains orientation sensitivity for elongated structures. Optimal performance was achieved with $\omega_1 = 0.6$ and $\omega_2 = 0.4$, resulting in mAP of 84.9 %, representing a 1.6 % improvement over the baseline model. This configuration demonstrates GeoCIoU's optimal optimization capability. Based on these results, we adopted $\omega_1 = 0.6$ and $\omega_2 = 0.4$ for all subsequent experiments.

**TABLE 2. Experimental results with different weights of GeoCIoU.**

| Methods | $\omega_1$ | $\omega_2$ | AP % | $mAP_{0.5}$ % |
|---|---|---|---|---|
| **Baseline (CIoU)** | / | / | 83.2 | 84.5 |
| **+GeoCIoU** | 0.5 | 0.5 | 83.9 | 84.8 |
| **+GeoCIoU** | **0.6** | **0.4** | **84.8** | **84.9** |
| **+GeoCIoU** | 0.7 | 0.3 | 84.5 | 84.5 |

### E. COMPARISON EXPERIMENT

To validate the effectiveness of the proposed model, a series of experiments are conducted on the NWPU VHR-10 and VisDrone datasets, comparing SWL-YOLO with mainstream object detection models, including Faster R-CNN, YOLOv5, YOLOv8, YOLOv10, Mask R-CNN, YOLOP, YOLOv11, as well as ADS-YOLO [40], CANet [41], SCRDet [42], and CE-FPN [43].

For the object detection subtask, SWL-YOLO achieved the best overall performance among all compared models. Evaluation was based on $mAP_{0.5}$, with results summarized in TABLE 3. Our method obtains higher mAP scores in categories such as airplanes, ships, storage tank, tennis court, ground track field, harbors, and bridges. SWL-YOLO outperformed both two-stage algorithms, Faster R-CNN and Mask R-CNN, as well as one-stage models including S2A-Net [44] and SCRDet, demonstrating superior detection capability across diverse remote sensing object categories.

**TABLE 3. Comparison of different models on NWPU VHR-10 dataset (%).**

| Model | airplane | ship | tank | baseball | court | harbor | bridge | vehicle | runway | motor | $mAP_{0.5}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Faster-RCNN | 89.2 | 78.5 | 82.1 | 73.4 | 85.6 | 76.3 | 80.7 | 75.8 | 88.9 | 72.4 | 80.3 |
| YOLOv5 | 88.7 | 77.8 | 81.5 | 72.9 | 84.2 | 75.1 | 79.3 | 74.6 | 87.5 | 71.8 | 79.4 |
| YOLOv8 | 90.1 | 79.3 | 83.0 | 74.5 | 86.4 | 77.2 | 81.5 | 76.7 | 89.3 | 73.2 | 81.1 |
| YOLOv10 | 90.5 | 80.0 | 83.6 | 75.2 | 87.1 | 78.0 | 82.1 | 77.5 | 90 | 74 | 81.8 |
| Mask-RCNN | 89.8 | 79.1 | 82.7 | 71.8 | 86.0 | 76.8 | 81.0 | 76.2 | 89 | 73 | 80.8 |
| YOLOP | 88.0 | 76.5 | 80.2 | 71.8 | 83.5 | 74 | 78.5 | 73.5 | 86.8 | 70.5 | 78.3 |
| YOLO11s | 91.0 | 80.5 | 84.2 | 75.8 | 87.6 | 78.7 | 82.8 | 78 | 90.5 | 74.8 | 82.5 |
| ADS-YOLO | 91.3 | 81.0 | 84.7 | 76.2 | 88.0 | 79.2 | 83.3 | 78.5 | 91.0 | 75.3 | 83.0 |
| CANet | 92.0 | 82.1 | 85.5 | 77.0 | 88.0 | 80.0 | 84.0 | 79.3 | 91.8 | 76.2 | 83.9 |
| CE-FPN | 92.3 | 82.4 | 84.8 | 76.5 | 85.5 | 83.2 | 83.1 | 78.7 | 90.5 | 77.9 | 82.4 |
| SCRDet | 92.5 | 82.6 | 86.0 | 77.5 | 89.3 | 80.5 | 84.5 | 80.0 | 92.3 | 76.8 | 84.4 |
| Ours | 93.2 | 83.5 | 83.5 | 78.3 | 90.0 | 81.2 | 85.2 | 80.8 | 93.0 | 77.5 | 86.9 |

To comprehensively evaluate the effectiveness of the improved models for target detection, several current mainstream models were selected for comparation on the VisDrone dataset. SWL-YOLO was compared with Mask-RCNN, YOLOV5, YOLOV8, FCOS [45], M2Det [46], Zhang et al. [47], Xue et al. [48], Bai et al. [49], Drone-DETR [50] and YOLO11s. The experimental findings are presented in TABLE 4. Upon analyzing the results, the $mAP_{0.5}$ value of this model is particularly outstanding, as high as 0.65, which is higher than that of all the compared models. As a comprehensive metric of precision and recall, $mAP_{0.5}$ indicates that our model achieves an excellent balance between detection accuracy and robustness. The parameter count of our approach is 8.36M, with GFLOPs of 25.4, comparable to commonly used lightweight models. While maintaining low parameter count and computational requirements, the improved model in this paper demonstrates the highest performance, which strongly validates the effectiveness of the proposed strategy. While transformer-based detectors such as Drone-DETR have achieved strong global reasoning, their computational complexity limits deployment in real-time remote sensing scenarios. In contrast, SWL-YOLO leverages lightweight convolutional and wavelet-based mechanisms to achieve comparable performance with significantly lower resource requirements.

**TABLE 4. Comparison of different models on VisDrone dataset (%).**

| Model | AP/ % | | | | | $mAP_{0.5}$ | #params | GFLOPs |
|---|---|---|---|---|---|---|---|---|
| | person | car | van | truck | motor | | | |
| Mask-RCNN | 61.2 | 67.5 | 72.4 | 60.9 | 72.5 | 62.7 | 84.6 | 38.6 |
| YOLOV5 | 59.4 | 58.6 | 58.3 | 60.6 | 59.1 | 57.2 | 8.24 | 31.4 |
| YOLOV8 | 56.4 | 59.7 | 62.1 | 58.6 | 59.5 | 58.8 | 9.67 | 27.5 |
| FCOS | 57.8 | 56.7 | 58.5 | 53.2 | 52.4 | 54.5 | 36.9 | 28.9 |
| M2Det | 68.1 | 57.8 | 54 | 68.9 | 60.2 | 61.7 | 86.5 | 33.5 |
| Zhang et al. | 62.3 | 59.8 | 61.5 | 60.7. | 61.4 | 60.3 | 14.6 | 27.8 |
| Xue et al. | 60.5 | 64.6 | 62 | 65.7 | 63.2 | 62.4 | 8.94 | 35.7 |
| Bai et al. | 62.8 | 61.5 | 60.4 | 63.8 | 60.7 | 61.5 | 9.43 | 26.3 |
| Drone-DETR | 57.8 | 58.6 | 52.1 | 53.5 | 51.3 | 53.9 | 28.7 | 128.3 |
| YOLOv11 | 63.5 | 64.2 | 58.4 | 56.6 | 59.7 | 59.9 | 9.41 | 21.3 |
| ours | 78.2 | 82.4 | 76.3 | 61.9 | 75.8 | 65 | 8.36 | 25.4 |

### F. VISUALIZATION

To validate the performance of the SWL-YOLO algorithm for RSOD, representative examples were selected from the NWPU VHR-10 and VisDrone datasets. We carried out a comparison of the benchmark algorithm YOLOv11 and the improved SWL-YOLO. FIGURE 10 presents the visualization results of both models. Additionally, heatmap-based visualizations of the model outputs are provided to intuitively demonstrate the advantages of the improved architecture in feature extraction, highlighting attention regions and feature responses to the input images.

As illustrated in FIGURE 10, the first visual comparison reveals that YOLOv11 suffers from false detection issues, whereas SWL-YOLO achieves more accurate pixel-level predictions. In the second comparison, SWL-YOLO demonstrates higher detection precision than YOLO11s. The third comparison shows that YOLO11s misidentifies distant vehicles, while SWL-YOLO maintains smoother and more complete segmentation edges. In the fourth

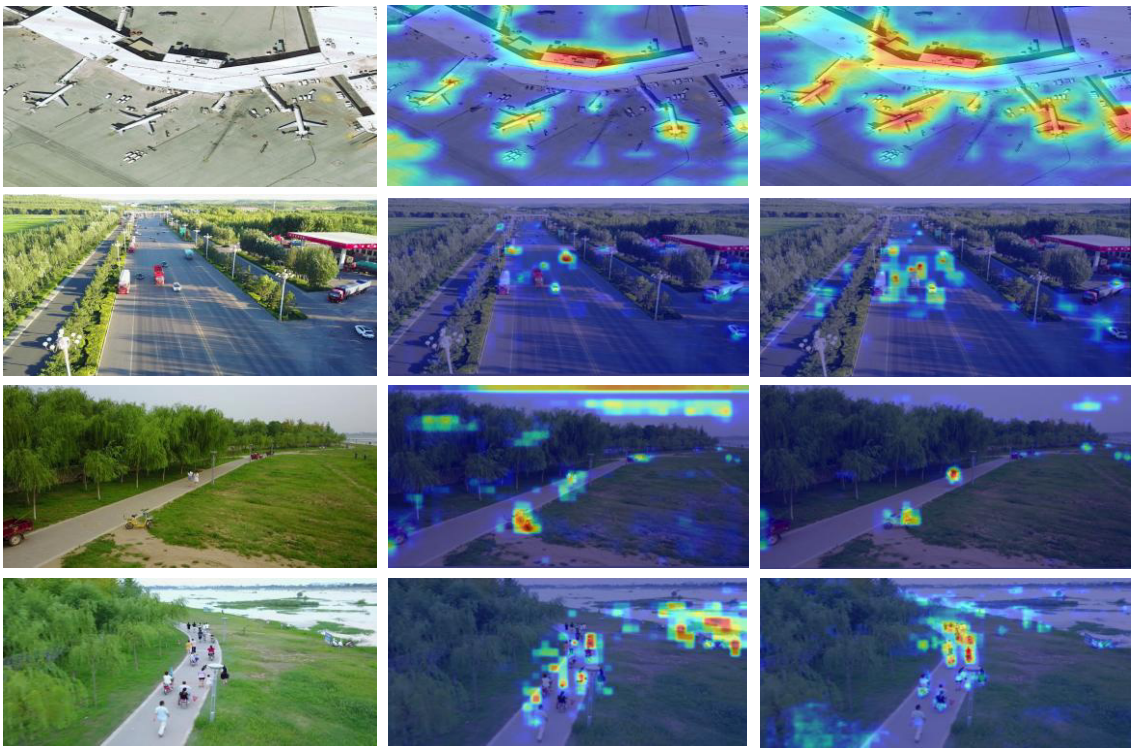
(a) GT images (b)YOLO11s (c)SWL-YOLO

**FIGURE 10.** **The visualization results on two datasets. While the baseline detector YOLO11s falsely detects or misses the small targets, the proposed SWL-YOLO detector can correctly detect the target objects.**

comparison, YOLO11s fails to detect certain objects, whereas SWL- YOLO successfully identifies distant motorcycles and pedestrians. In the fifth comparison, YOLO11s misclassifies a motorcycle as a pedestrian, while SWL-YOLO accurately detects all objects. The visualization results demonstrate that SWL-YOLO exhibits strong robustness and high precision, delivering reliable detection performance across diverse scenarios. This makes the model particularly well-suited for small object detection tasks in complex and challenging remote sensing environments.

Figure 11 illustrates the heatmap visualizations of both models on the VisDrone dataset. FIGURE 11 (a) shows the original image, while Figure 11 (b) and 11 (c) present the activation maps of YOLO11s and SWL-YOLO under identical conditions, respectively. In the heatmaps, red regions correspond to areas contributing most significantly to target detection. In Figure 11(b), YOLO11s exhibits diffused activation patterns with weaker responses to small targets, consistent with its lower recall (89.2% vs. 89.8% for SWL-YOLO as reported in Table 1). By contrast, SWL-YOLO's heatmap in Figure 11(c) demonstrates three key advantages: (1) Sharper high-frequency activations precisely aligned with small object edges (red regions), validating the effectiveness of the WDown module in preserving fine-grained spatial details via wavelet decomposition. (2) Stronger contextual awareness around clustered objects (blue-green halos), attributable to the LSK module's dynamic receptive field adjustment. (3) Suppressed background noise (fewer scattered activations), reflecting the SAFM's spatially-adaptive feature. weighting. These activation patterns correlate directly with SWL-YOLO's 5.1% $mAP_{0.5}$ improvement on the VisDrone dataset, particularly for sub-20px targets where edge and texture features dominate detection.

Compared to YOLO11s, SWL-YOLO not only accurately localizes critical features of the target but also effectively incorporates contextual background information, achieving more comprehensive feature fusion. This discrepancy indicates that the improved SWL-YOLO exhibits



(a)Images (b)YOLO11s (c)SWL-YOLO

**FIGURE 11.** Heat map visualization of baseline model and improved model on VisDrone dataset. In each subfigure, the first column shows the example images from VisDrone dataset. The second column shows their feature maps generated by YOLO11s, while the third column shows the feature maps generated by SWL-YOLO.

significant advantages in leveraging contextual information, thereby enhancing detection accuracy and semantic scene understanding. The heatmap visualization demonstrates that synergistic integration of SAFM (local feature module), WDown (detail preservation), and LSK (global context) provides stronger robustness and adaptive capability in feature extraction and background fusion. These results further validate the effectiveness of the proposed model under complex remote sensing scenarios.

## IV. CONCLUSION

In this study, we propose an enhanced object detection algorithm, SWL-YOLO, specifically designed to improve small object detection in low-quality RSIs. The model is built upon the YOLOv11 framework. To address the challenges of high computational overhead and limited detection accuracy for small-scale targets, several key enhancements are introduced: firstly, SAFM is introduced. By employing channel splitting and spatially-aware dynamic weighting, SAFM modulates feature maps at both region and pixel levels, improving the model's sensitivity to small objects and edge features while maintaining low computational overhead. Second, to tackle the challenges posed by small objects that rely on contextual information and varying receptive field requirements across object categories, the LSK mechanism is incorporated. LSK dynamically adjusts receptive fields through efficient decomposition of large kernels, enabling adaptive contextual modeling that substantially enhances detection robustness. Third, to mitigate issues related to low contrast and loss of detail in degraded RSIs, the WDown operator is adopted. This operator effectively preserves high-frequency details without blurring or discarding critical spatial features. Finally, for improved bounding box regression, the original CIoU loss is replaced with the GeoCIoU loss, which employs a dual-penalty mechanism to provide more accurate training feedback regarding object orientation, scale, and overlap.

Experimental results demonstrate that SWL-YOLO achieves consistent improvements over YOLOv11, with $mAP_{0.5}$ gains of 5.1 % on VisDrone and 4.2 % on NWPU VHR-10, while reducing parameters by around 10 %. The model maintains real-time inference capability and strikes a practical balance between accuracy and efficiency, making it well-suited for time-sensitive remote sensing tasks. Despite the improvements, limitations remain when detecting ultra-small or densely overlapping objects and when using RGB-only imagery. Future work will explore multimodal fusion and transformer integration to enhance robustness.

## REFERENCES

[1] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[2] G. Priyadharshini and D. R. J. Dolly, "Comparative investigations on tomato leaf disease detection and classification using CNN, R-CNN, fast R-CNN and faster R-CNN," in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2023, pp. 1540–1545.
[3] H. Qin, J. Wang, X. Mao, Z. Zhao, X. Gao, and W. Lu, "An improved faster R-CNN method for landslide detection in remote sensing images," *J. Geovisualization Spatial Anal.*, vol. 8, no. 1, p. 2, Jun. 2024.
[4] R. Sapkota, D. Ahmed, and M. Karkee, "Comparing YOLOv8 and mask R-CNN for instance segmentation in complex orchard environments," *Artif. Intell. Agricult.*, vol. 13, pp. 84–99, Sep. 2024.
[5] O. E. Olorunshola, M. E. Irhebhude, and A. E. Evwiekpaefe, "A comparative study of YOLOv5 and YOLOv7 object detection algorithms," *J. Comput. Social Informat.*, vol. 2, no. 1, pp. 1–12, Feb. 2023.
[6] F. M. Talaat and H. ZainEldin, "An improved fire detection approach based on YOLO-v8 for smart cities," *Neural Comput. Appl.*, vol. 35, no. 28, pp. 20939–20954, 2023.
[7] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Mar. 2024, pp. 107984–108011.
[8] J. Zhan, Y. Luo, C. Guo, Y. Wu, J. Meng, and J. Liu, "YOLOPX: Anchor-free multi-task learning network for panoptic driving perception," *Pattern Recognit.*, vol. 148, Apr. 2024, Art. no. 110152.
[9] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.
[10] R. Mahum and A. S. Al-Salman, "Lung-RetinaNet: Lung cancer detection using a RetinaNet with multi-scale feature fusion and context module," *IEEE Access*, vol. 11, pp. 53850–53861, 2023.
[11] H. Zhao, Y. Gao, and W. Deng, "Defect detection using shuffle Net-CA-SSD lightweight network for turbine blades in IoT," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 32804–32812, Oct. 2024.
[12] Z. Li, J. Yuan, G. Li, H. Wang, X. Li, D. Li, and X. Wang, "RSI-YOLO: Object detection method for remote sensing images based on improved YOLO," *Sensors*, vol. 23, no. 14, pp. 6414–6427, Jul. 2023.
[13] M. Cui et al., "LC-YOLO: A lightweight model for small object detection," *Appl. Sci.*, vol. 13, no. 5, pp. 3174–3190, 2023.
[14] R. Zhang, C. Xie, and L. Deng, "A fine-grained object detection model for aerial images based on YOLOv5 deep neural network," *Chin. J. Electron.*, vol. 32, no. 1, pp. 51–63, Jan. 2023.
[15] J. Zhang, Z. Chen, G. Yan, Y. Wang, and B. Hu, "Faster and lightweight: An improved YOLOv5 object detector for remote sensing images," *Remote Sens.*, vol. 15, no. 20, pp. 4974–4990, Oct. 2023.
[16] J. Yi, Z. Shen, F. Chen, Y. Zhao, S. Xiao, and W. Zhou, "A lightweight multiscale feature fusion network for remote sensing object counting," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.
[17] S. Liang, H. Wu, L. Zhen, Q. Hua, S. Garg, G. Kaddoum, M. M. Hassan, and K. Yu, "Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25345–25360, Dec. 2022.
[18] L. Y. Wu, L. Liu, Y. Wang, Z. Zhang, F. Boussaid, M. Bennamoun, and X. Xie, "Learning resolution-adaptive representations for cross-resolution person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 4800–4811, 2023.
[19] Y. Li, J. Wu, W. Chen, P. Tan, C.-T. Ngan, and B. Ou, "A lightweight and multi-branch module in facial semantic segmentation feature extraction," *IEEE Access*, vol. 12, pp. 84803–84814, 2024.
[20] Z. Su, J. Yu, H. Tan, X. Wan, and K. Qi, "MSA-YOLO: A remote sensing object detection model based on multi-scale strip attention," *Sensors*, vol. 23, no. 15, pp. 6811–6825, Jul. 2023.
[21] L. Zhou, Y. Li, X. Rao, C. Liu, X. Zuo, and Y. Liu, "Ship target detection in optical remote sensing images based on multiscale feature enhancement," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–20, Oct. 2022.
[22] T. Gao, Z. Liu, J. Zhang, G. Wu, and T. Chen, "A task-balanced multiscale adaptive fusion network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023.
[23] H. Liao and W. Zhu, "YOLO-DRS: A bioinspired object detection algorithm for remote sensing images incorporating a multi-scale efficient lightweight attention mechanism," *Biomimetics*, vol. 8, no. 6, p. 458, Oct. 2023.
[24] Y. Wu et al., "AMR-Net: Arbitrary-oriented ship detection using attention module," *IEEE Access*, vol. 9, pp. 68208–68222, 2021.
[25] G. Li, Q. Fang, L. Zha, X. Gao, and N. Zheng, "HAM: Hybrid attention module in deep convolutional neural networks for image classification," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108785.
[26] H. Xu, "Rotated single-stage detector with transformer in remote sensing rotating object detection," in *Proc. 5th Int. Conf. Comput. Eng. Appl. (ICCEA)*, Apr. 2024, pp. 1298–1303.
[27] S. Mei, "Rotation-invariant feature learning via convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600713.
[28] L. Yang, H. Wang, Y. Fang, S. Wang, L. Zhi, and B. Jiang, "Learning power Gaussian modeling loss for dense rotated object detection in remote sensing images," *Chin. J. Aeronaut.*, vol. 36, no. 10, pp. 353–365, Oct. 2023.
[29] P. Huang, S. Tian, Y. Su, W. Tan, Y. Dong, and W. Xu, "IA-CIOU: An improved IOU bounding box loss function for SAR ship target detection methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 10569–10582, 2024.
[30] N. Jegham, C. Y. Koh, M. Abdelatti, and A. Hendawi, "YOLO evolution: A comprehensive benchmark and architectural review of YOLOv12, YOLO11, and their previous versions," 2024, *arXiv:2411.00201*.
[31] J. Zhao, S. Miao, R. Kang, L. Cao, L. Zhang, and Y. Ren, "Insulator defect detection algorithm based on improved YOLOv11n," *Sensors*, vol. 25, no. 5, p. 1327, Feb. 2025.
[32] H. K. Jooshin, M. Nangir, and H. Seyedarabi, "Iarnception-YOLO: Computational cost and accuracy improvement of YOLOv5," *IET Image Process.*, vol. 18, no. 8, pp. 1985–1999, 2024.
[33] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "HiFuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105534.
[34] M. Lee, "GELU activation function in deep learning: A comprehensive mathematical analysis and performance," 2023, *arXiv:2305.12073*.
[35] Q. Li, L. Shen, S. Guo, and Z. Lai, "Wavelet integrated CNNs for noise-robust image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7243–7252.
[36] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11975–11986.
[37] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16794–16805.
[38] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Feb. 2019.
[39] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310–314, Feb. 2019.
[40] M. Chen, M. Li, Q. Wang, and X. Cao, "ADS-YOLO: An enhanced YOLO framework for high-speed MLCCs defect detection," *Infr. Phys. Technol.*, vol. 145, Mar. 2025, Art. no. 105733.
[41] X. Hou, M. Liu, S. Zhang, P. Wei, and B. Chen, "CANet: Contextual information and spatial attention based network for detecting small defects in manufacturing industry," *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109558.
[42] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.
[43] Y. Luo, X. Cao, J. Zhang, J. Guo, H. Shen, T. Wang, and Q. Feng, "CE-FPN: Enhancing channel information for object detection," *Multimedia Tools Appl.*, vol. 81, no. 21, pp. 30685–30704, Sep. 2022.
[44] J. Li, M. Chen, S. Hou, Y. Wang, Q. Luo, and C. Wang, "An improved S2A-Net algorithm for ship object detection in optical remote sensing images," *Remote Sens.*, vol. 15, no. 18, p. 4559, Sep. 2023.
[45] R. Zhao, Y. Guan, Y. Lu, Z. Ji, X. Yin, and W. Jia, "FCOS-LSC: A novel model for green fruit detection in a complex orchard environment," *Plant Phenomics*, vol. 5, p. 69, May 2023.
[46] X. Zhang, Z. Gong, H. Guo, X. Liu, L. Ding, K. Zhu, and J. Wang, "Adaptive adjacent layer feature fusion for object detection in remote sensing images," *Remote Sens.*, vol. 15, no. 17, p. 4224, Aug. 2023.
[47] N. Wang, Y. Gao, H. Chen, P. Wang, Z. Tian, C. Shen, and Y. Zhang, "NAS-FCOS: Efficient search for object detection architectures," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3299–3312, Dec. 2021.
[48] J. Xue, Y. Zheng, C. Dong-Ye, P. Wang, and M. Yasir, "Improved YOLOv5 network method for remote sensing image-based ground objects recognition," *Soft Comput.*, vol. 26, no. 20, pp. 10879–10889, Oct. 2022, doi: 10.1007/s00500-022-07106-8.

[49] X. Bai, X. Li, J. Miao, and H. Shen, "A front-back view fusion strategy and a novel dataset for super tiny object detection in remote sensing imagery," *Knowl.-Based Syst.*, vol. 326, Sep. 2025, Art. no. 114051, doi: 10.1016/j.knosys.2025.114051.
[50] Y. Kong, X. Shang, and S. Jia, "Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model," *Sensors*, vol. 24, no. 17, p. 5496, Aug. 2024, doi: 10.3390/s24175496.

**JING ZHANG** received the master's degree in computational mathematics from Dalian University of Technology, in 2016. She is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She is also a Lecturer with the Public Teaching Department, Anhui Vocational College of City Management. Her research interests include computer vision and object detection.

**MAS RINA BINTI MUSTAFFA** (Senior Member, IEEE) received the B.Comp.Sc. degree in multimedia and the M.Sc. and Ph.D. degrees in multimedia systems from Universiti Putra Malaysia (UPM), Malaysia, in 2003, 2006, and 2012, respectively.

She is currently an Associate Professor with the Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She has authored numerous indexed publications and led various research projects. Her current research interests include computer vision, multimedia analytics, pattern recognition, multimedia information retrieval, and deep learning for image analysis.

Dr. Mustaffa is a registered Professional Technologist (Ts.). She served on the organizing committees for several IEEE-sponsored international conferences and regularly contributes as a reviewer for journal and conference publications.

**FATIMAH KHALID** (Member, IEEE) received the B.Sc. degree in computer science from the University of Technology Malaysia (UTM), in 1992, and the master's degree and the Ph.D. degree in system science and management from Universiti Kebangsaan Malaysia (UKM), in 1997 and 2008, respectively. From 1993 to 1995, she was a System Analyst with UKM. After the master's degree, she started involved in teaching with the Sal College, until 1999, and continued with Universiti Putra Malaysia, in June 1999. In January 2016, she was on a Secondment with the Computer Science Department, Tabuk University, Saudi Arabia, for one and a half years. She is currently with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, as a Lecturer and an Associate Professor.

**ZAINAL ABDUL KAHAR** (Member, IEEE) received the B.Sc. degree from the University of Technology Malaysia (UTM), and the M.Sc. and Ph.D. degrees from Universiti Putra Malaysia (UPM). He is currently an Associate Professor with the Department of Multimedia, Faculty of Computer Science and Information Technology, University Putra Malaysia. His expertise spans game design, computer graphics, and image processing, with a focus on innovative applications in interactive media, 3D rendering, and machine learning-based visual analysis.