Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082486

ARTICLE

Dual-Strategy Improvement of YOLOv11n for Multi-Scale Object Detection in Remote Sensing Images

Shuaiyu Zhu[1], **Sergey Ablameyko**[1,2] **and Ji Li**[3,*]

[1]Faculty of Mechanics and Mathematics, Belarusian State University, Minsk, Belarus
[2]United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus
[3]School of Computer Science and Engineering, Northeastern University, Shenyang, China

*Corresponding Author: Ji Li. Email: jilisch@163.com

Received: 17 March 2026; Accepted: 15 April 2026; Published: 15 June 2026

ABSTRACT: Satellite remote sensing images pose significant challenges for object detection due to their high resolution, complex scenes, and large variations in target scales. To address the insufficient detection accuracy of the YOLOv11n model in remote sensing imagery, this paper proposes two improvement strategies. Method 1: (a) a Large Separable Kernel Attention (LSKA) mechanism is introduced into the backbone network to enhance feature extraction for small objects; (b) a Gold-YOLO structure is incorporated into the neck network to achieve multi-scale feature fusion, thereby improving the detection performance of objects at different scales. Method 2: (a) the Gold-YOLO structure is also integrated into the neck network; (b) a MultiSEAMHead detection head is combined to further strengthen the representation and detection capability for small and multi-scale objects. To verify the effectiveness of the proposed improvements, experiments are conducted on the DOTAv1 dataset. The results show that, while maintaining the lightweight advantage of the model, the proposed methods improve detection accuracy (mAP@0.5) by 1.3% and 1.8%, respectively, compared with the baseline YOLOv11n, demonstrating the effectiveness and practical value of the proposed approaches for object detection in remote sensing images.

KEYWORDS: Remote sensing imagery; YOLOv11n; multi-scale object detection; lightweight deep learning; attention mechanism; feature fusion

1 Introduction

With the continuous advancement of remote sensing imaging technologies and deep learning algorithms, deep learning–based object detection in remote sensing images has demonstrated significant application potential in fields such as urban governance, smart agriculture, and national defense security [1]. However, in real-world complex scenarios, this technology still faces multiple challenges. First, due to the long imaging distance of satellite and aerial platforms, targets in high-resolution remote sensing images usually occupy only a very small number of pixels (most are smaller than 32 × 32 pixels), which makes effective feature extraction extremely difficult. Second, remote sensing images are characterized by complex and diverse background structures, where urban buildings, transportation facilities, farmland, and natural terrain are interwoven. This complexity reduces the contrast between targets and the background and easily leads to feature confusion. Third, remote sensing scenes contain a wide variety of object categories with significant appearance variations. Even within the same category, substantial differences in scale, shape, and texture may exist, further increasing the difficulty of model generalization across diverse scenarios.

In addition to small object detection, multi-scale variation is one of the most fundamental challenges in remote sensing imagery. Objects in aerial images exhibit significant scale differences, ranging from vehicles that occupy only a few pixels to large buildings spanning hundreds of pixels, resulting in a substantial scale span. This wide variation makes it difficult for traditional feature pyramid structures to simultaneously preserve fine-grained spatial details and high-level semantic information. Therefore, enhancing cross-scale feature interaction and adaptive multi-level fusion is crucial for achieving robust multi-scale object detection.

Although detection performance has been continuously improved in recent years through advances in network architectures, feature enhancement strategies, and optimized training mechanisms, issues such as missed detection of small objects and false detections in complex backgrounds remain prominent due to the limited feature representation capability [2]. Therefore, research efforts focusing on enhancing feature representation, multi-scale feature fusion, and improving model robustness have become key directions for further advancing remote sensing object detection technologies [3].

Deep learning–based object detection methods have evolved from early two-stage approaches to lightweight, high-speed, and Transformer-driven architectures [4]. In practical applications, researchers typically select appropriate models according to specific scenario requirements: two-stage detectors are preferred when high accuracy is required; single-stage detectors such as YOLO [5] and SSD [6] are more commonly adopted when real-time performance is prioritized; and in scenarios with complex backgrounds and pronounced long-range dependencies, Transformer-based detectors are increasingly demonstrating greater potential.

In response to these challenges, researchers have explored various strategies to enhance feature representation and cross-scale information interaction in remote sensing object detection. Xu and Mao [7] explored multi-level feature fusion for UAV aerial-image detection. Du et al. [8] proposed CFPT to enhance cross-layer interaction and reduce semantic gaps in aerial-image small object detection. Zhou et al. [9] developed SMA-YOLO with sparse attention, a bidirectional auxiliary feature pyramid, and an adaptive head to improve feature aggregation and robustness in UAV remote sensing scenes. Song et al. [10] further improved fine-grained remote sensing detection through a refined-balanced feature pyramid network and a center-scale-aware label assignment strategy. Recent research has mainly focused on improving multi-scale feature fusion and small target detection in remote sensing imagery. Yuan et al. [11] proposed using shallow features to provide structural anchors for deep features, injecting lost high-frequency information, and achieving adaptive balance through a cascaded gating fusion mechanism, which effectively suppresses background noise and enhances small target responses, thereby improving multi-scale fusion performance. Li and Qu [12] introduced a feature distribution technique based on multi-scale feature fusion strategies to achieve more efficient cross-scale interaction. Fu et al. [13] designed a differential enhancer that combines statistical differences and structural similarity to guide the model toward more discriminative features, while also strengthening edge information through an edge-enhanced selection perceptron to boost small target detection. Liu and Li [14] introduced the normalized Wasserstein distance in mixed-loss training to better emphasize small target features and suppress background interference. Although these studies have made significant progress in addressing challenges such as object diversity and complex backgrounds in remote sensing images, there remains considerable room for improvement in handling weak target features and achieving higher detection accuracy for small-sized objects.

To further quantify the limitations of YOLOv11n in remote sensing tasks, we conducted preliminary experiments on the DOTA-v1 dataset. YOLOv11n achieves a mAP50 of 42.0% and a mAP50–95 of 25.7%. However, its performance on small-scale categories remains limited, with an AP of only 53.9% for small vehicles and an AP below 37% for ships, indicating its insufficient ability to capture fine-grained features. The recall of 40.2% suggests that a considerable number of objects are still missed, especially in object-dense or

small-object scenes. These results demonstrate that YOLOv11n suffers from inadequate multi-scale feature representation and weak sensitivity to small objects in remote sensing imagery.

To address the aforementioned issues, this paper proposes two improved lightweight object detection models based on YOLOv11n. Compared with existing methods, the main contributions of this paper are as follows: a dual-strategy lightweight improvement framework is proposed to overcome the limitations of YOLOv11n in feature extraction and feature fusion; the LSKA module is introduced to enhance large receptive field perception with minimal computational cost, specifically for enhancing small target features in remote sensing scenes [15]; a Gold-YOLO-based neck structure is integrated to improve cross-scale feature interaction through an aggregation-distribution mechanism; a systematic analysis of the MultiSEAMHead is conducted, revealing the dependence of its performance on the neck structure; and a comparative study of two optimized propeller heads is performed to guide practical deployment under different constraints. Extensive experiments conducted on the DOTA-v1 dataset demonstrate the effectiveness and practical value of the proposed approaches.

2 Model Frame

2.1 YOLOv11n Model

YOLOv11n [16] is a lightweight single-stage object detector designed to balance detection accuracy and computational efficiency. Compared with earlier YOLOv8 [17] versions, it introduces improved feature extraction and lightweight design strategies, making it suitable for real-time applications. However, when applied to high-resolution remote sensing images, YOLOv11n still exhibits limitations in global contextual modeling and cross-scale feature interaction, particularly for small objects and densely distributed targets. These limitations motivate further architectural enhancements tailored to remote sensing scenarios. Fig. 1 illustrates the architecture of YOLOv11n.

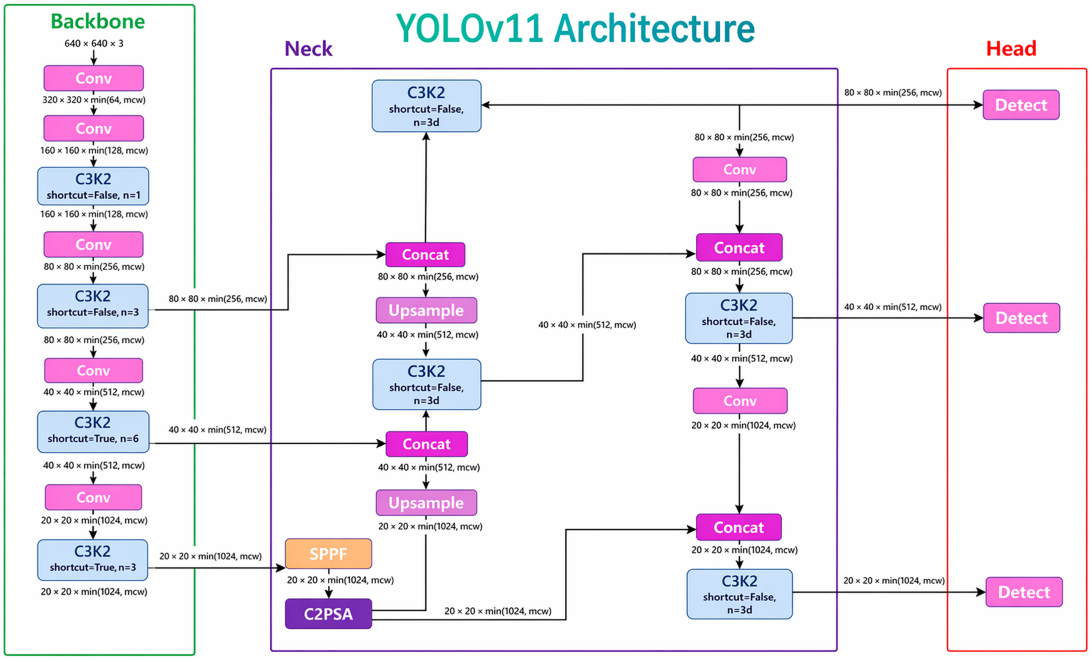


Figure 1: YOLOv11 architecture [18].

2.2 GOLD-YOLO

Gold-YOLO [19] is an efficient object detection framework designed to improve multi-scale feature fusion through a gather-and-distribute (GD) mechanism [20]. By jointly aggregating features from different network depths and redistributing the fused information across layers, Gold-YOLO enhances cross-scale information flow while maintaining low latency and computational efficiency. This design is particularly beneficial for object detection tasks involving large scale variations and dense object distributions, which are common in remote sensing imagery.

As illustrated in Fig. 2, the Gold-YOLO architecture consists of a backbone network, a GD-based neck, and a detection head. The GD structure is composed of two complementary branches: a low-level gather-and-distribute (Low-GD) branch and a high-level gather-and-distribute (High-GD) branch. The Low-GD branch focuses on processing shallow, high-resolution feature maps to preserve fine-grained spatial details that are critical for small object detection, while the High-GD branch emphasizes deep semantic features to enhance contextual understanding for medium and large objects.

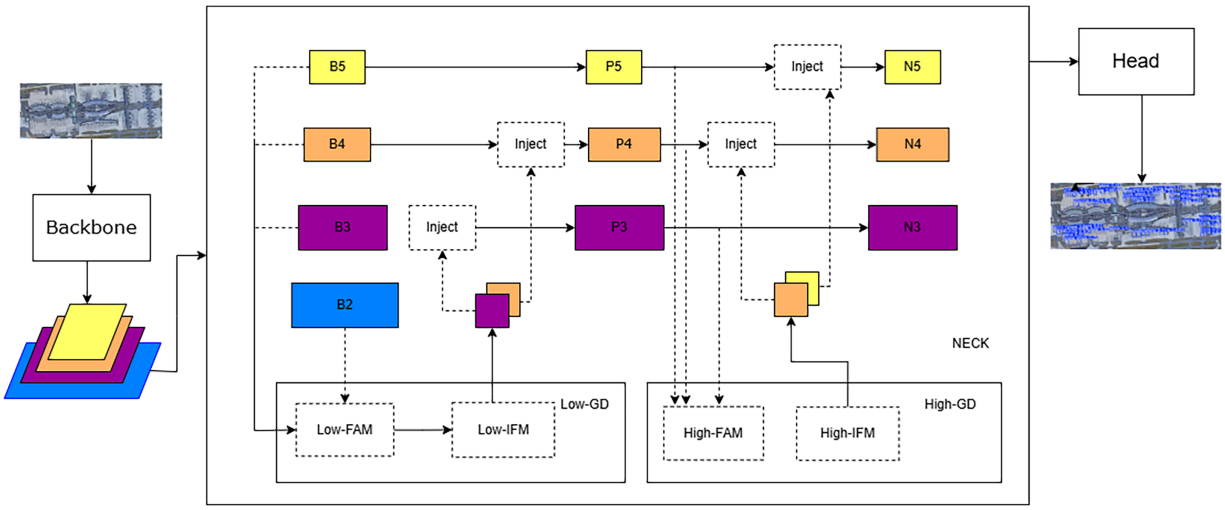


Figure 2: Gold-YOLO architecture.

Within each branch, feature alignment and information fusion are performed to alleviate feature misalignment caused by scale differences. Specifically, feature alignment modules (FAM) are employed to spatially align features from adjacent layers, followed by information fusion modules (IFM) that integrate aligned features to generate more consistent representations [21]. The fused features are then injected back into different feature levels through the Inject module, enabling effective information redistribution and enhancement across the network.

In addition, Gold-YOLO incorporates a Lightweight Adjacent Fusion (LAF) module to further strengthen feature interaction between neighboring layers [22]. By performing scale alignment via pooling and up/down-sampling and adopting lightweight fusion operations, LAF allows each feature level to receive complementary information from adjacent layers with minimal computational overhead. Through the coordinated operation of the GD mechanism and LAF module, Gold-YOLO achieves more robust and consistent multi-scale feature representations, providing a solid foundation for improving detection performance in complex remote sensing scenes.

In remote sensing scenarios, objects are often small and surrounded by complex backgrounds. The large receptive field provided by LSKA enables the model to capture broader contextual cues, which is essential

for distinguishing small objects from cluttered backgrounds. Therefore, LSKA is not only efficient but also particularly suitable for remote sensing tasks.

2.3 Large Separable Kernel Attention Mechanism

Large Separable Kernel Attention (LSKA) is an efficient attention mechanism designed to capture large receptive fields while maintaining low computational complexity [23]. In remote sensing images, objects are often surrounded by complex backgrounds and exhibit weak visual contrast, making global contextual information particularly important for accurate detection. However, directly employing large-kernel convolutions leads to a significant increase in parameters and computational cost, which is unsuitable for lightweight detection models.

LSKA addresses this issue by decomposing large two-dimensional convolution kernels into a series of separable one-dimensional convolutions along the horizontal and vertical directions, combined with depthwise and dilated convolution operations. This design enables the network to approximate large receptive fields efficiently while significantly reducing parameter count and computational overhead. Compared with conventional large-kernel attention mechanisms, LSKA achieves comparable global feature modeling capability with improved efficiency and stability [24].

By generating spatial attention maps based on large-context perception and reweighting the original feature maps, LSKA enhances the network's ability to focus on salient regions and suppress background interference. Owing to its favorable balance between global contextual modeling and lightweight design, LSKA is well suited for integration into YOLOv11n to improve feature extraction performance for small objects and dense scenes in remote sensing imagery.

The gather-and-distribute mechanism of Gold-YOLO aligns well with the multi-scale characteristics of remote sensing images, where objects of drastically different sizes coexist. Its ability to redistribute fused features enhances both small-object detail preservation and large-object semantic understanding.

2.4 MultiSEAMHead

Although the original YOLOv11n detection head achieves high inference efficiency, its ability to jointly exploit multi-level semantic and fine-grained features remains limited. Shallow features often lack sufficient semantic information, while deep features tend to lose spatial details, which negatively affects detection performance in complex remote sensing scenes characterized by dense object distributions, occlusion, and background clutter.

To address this limitation, the MultiSEAMHead (Fig. 3) module enhances the detection head by incorporating multi-level feature fusion and attention-based feature modulation [25]. It leverages depthwise separable convolutions and cross-layer connections to improve information interaction across different feature scales, enabling the detector to better utilize both detailed spatial information and high-level semantic representations. Specifically, the MultiSEAMHead consists of depthwise separable convolutional layers with kernel sizes of 3×3 and 5×5, followed by a channel mixing and spatial attention module. The number of channels matches the output feature maps of the neck. During integration, the original YOLOv11n detection head is replaced by the MultiSEAMHead. Input feature maps from the Gold-YOLO neck are fed into the MultiSEAMHead at three scales (P3, P4, P5), thereby achieving enhanced multi-level feature aggregation before classification and regression.

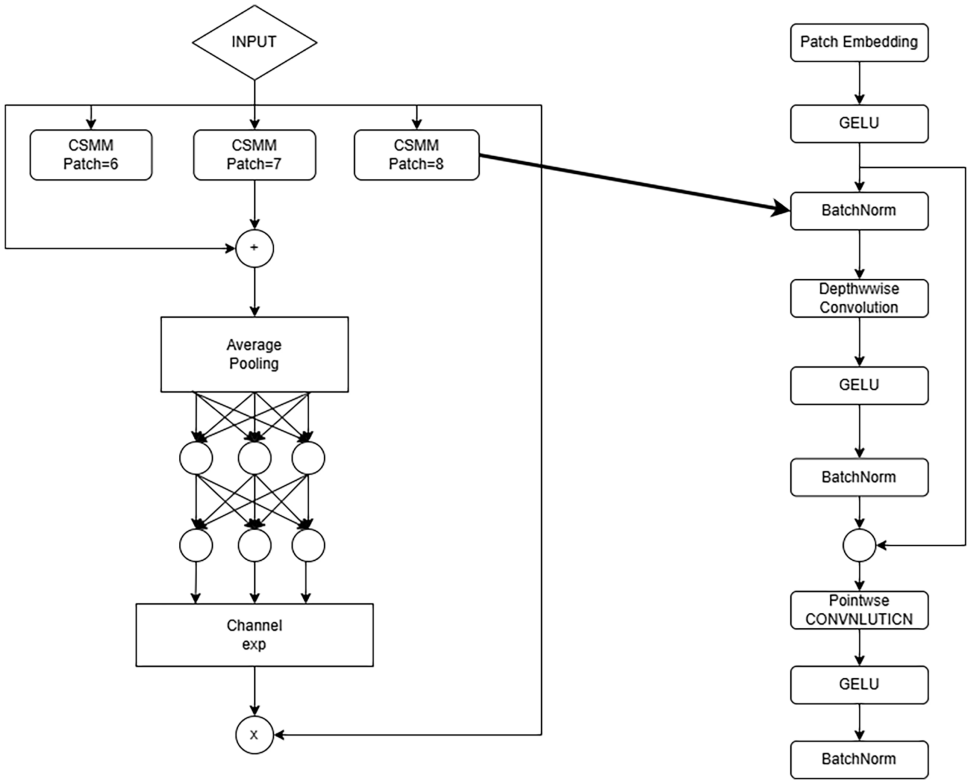


Figure 3: MultiSEAMHead structure diagram.

Furthermore, MultiSEAMHead introduces channel and spatial mixing mechanisms to model inter-channel dependencies and emphasize informative regions. By adaptively reweighting features through attention-based regulation, the detection head becomes more robust to scale variation and background interference. As a result, MultiSEAMHead serves as a candidate enhancement for handling small, occluded, or scale-sensitive objects, although its practical benefit depends on its interaction with the feature-fusion structure.

MultiSEAMHead enhances cross-layer feature interaction and attention modeling, which is beneficial for dense and occluded object detection commonly observed in remote sensing imagery.

3 Model Design

Although YOLOv11n introduces self-attention and lightweight convolutional designs, its architecture is primarily optimized for general-purpose object detection tasks. When applied to remote sensing images, several inherent limitations become apparent. First, the backbone network lacks sufficient large-receptive-field modeling capability, which restricts its ability to capture global contextual information crucial for distinguishing small objects from complex backgrounds. Second, the original neck structure provides limited cross-layer information interaction, leading to suboptimal feature fusion for objects with large scale variations. Finally, the detection head focuses mainly on efficiency, while the joint utilization of multi-level semantic and fine-grained features remains insufficient for robust detection in dense and cluttered scenes.

Motivated by these observations, we introduce targeted architectural modifications to YOLOv11n. The LSKA mechanism is employed to enhance global feature perception with minimal computational overhead, addressing the limited receptive field issue. The Gold-YOLO neck is integrated to strengthen

multi-scale feature aggregation and distribution, improving information flow across different feature levels. Furthermore, the MultiSEAMHead is adopted to enhance the detection head's ability to jointly model spatial and channel-wise dependencies. These modifications are specifically designed to overcome the challenges of remote sensing object detection while preserving the lightweight nature of YOLOv11n.

3.1 YOLOv11n-LSKA-GoldYOLO

First improved lightweight object detection model proposed in this study, YOLOv11n-LSKA-GoldYOLO, is built upon the YOLOv11n framework (Fig. 4). To enhance feature representation in complex scenes, a Large Separable Kernel Attention (LSKA) mechanism is introduced into the backbone network. LSKA employs a decomposed large convolution kernel structure that effectively enlarges the receptive field while keeping computational costs manageable, enabling the model to capture global contextual information more comprehensively and suppress redundant features. This design is intended to enhance contextual modeling and multi-scale feature propagation in complex remote sensing scenes, which may benefit certain small or densely distributed targets.

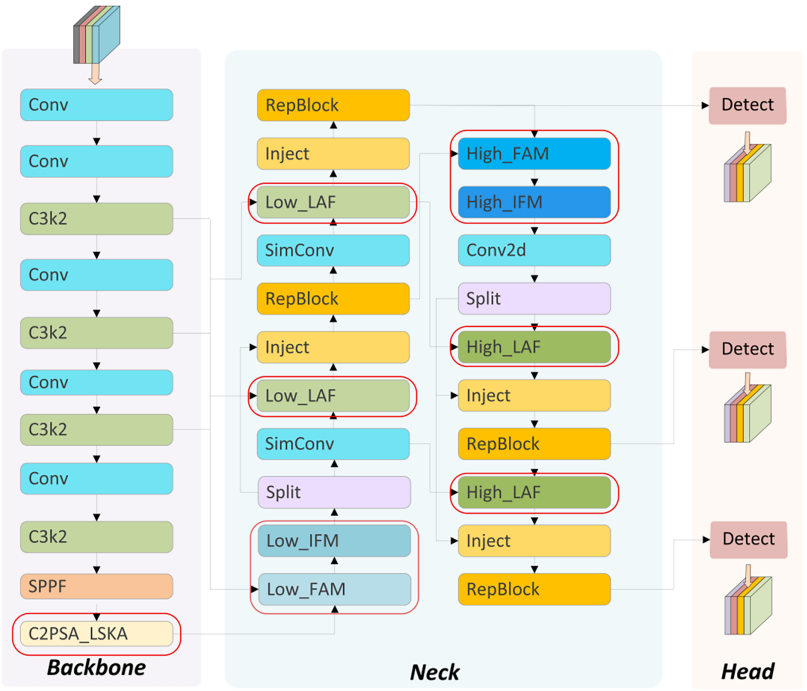


Figure 4: YOLOv11n-LSKA-GoldYOLO.

During the feature fusion stage, the Gold-YOLO Neck structure is further employed to achieve more efficient multi-scale information interaction and adaptive feature integration, thereby improving feature propagation and detection accuracy across objects of different scales. By combining the strengths of the LSKA mechanism and the Gold-YOLO Neck, YOLOv11n-LSKA-GoldYOLO significantly enhances detection performance while maintaining low computational cost and real-time inference capability, offering a practical and efficient solution for lightweight object detection tasks.

3.2 YOLOv11n-GoldYOLO-MultiSEAMHead

This study further proposes the lightweight object detection model YOLOv11n-GoldYOLO-MultiSEAMHead, whose overall architecture is based on YOLOv11n and incorporates the efficient feature fusion design of Gold-YOLO (Fig. 5). During the feature interaction stage, the model employs the Gold-YOLO Neck structure to facilitate more comprehensive multi-scale information flow and a more stable adaptive fusion mechanism, enabling efficient feature propagation between shallow and deep layers and improving detection performance across objects of varying scales.

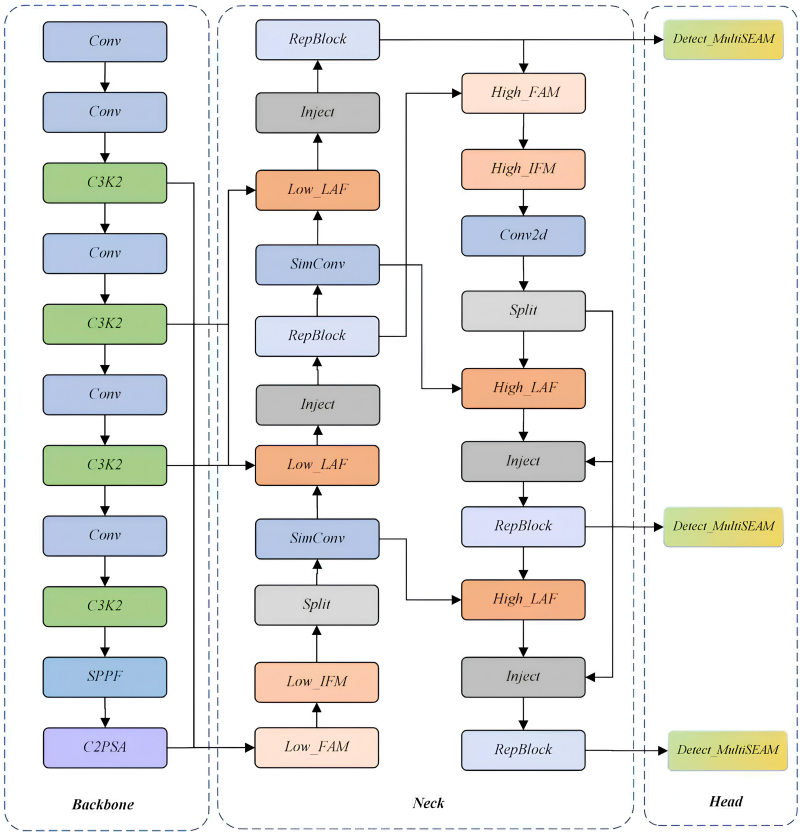


Figure 5: YOLOv11n-GoldYOLO-MultiSEAMHead.

To further enhance the representational capacity of the detection head, the model integrates the MultiSEAMHead module. This module performs joint modeling of multi-level features through depthwise separable convolutions, multi-scale feature extraction, and cross-layer connections, allowing the network to more comprehensively capture fine-grained details and high-level semantic information. By leveraging its built-in Channel and Spatial Mixing Units and globally context-aware attention weighting, MultiSEAMHead enables the network to accurately focus on key regions and suppress irrelevant features in scenarios involving occlusion, complex backgrounds, or dense objects. Additionally, the improved regression loss in the detection head further stabilizes the training process and accelerates the convergence of bounding box predictions.

4 Experimental Procedure

4.1 Experimental Dataset

The experiments in this study use the publicly available DOTA-v1 (Dataset for Object deTection in Aerial Images Version 1) dataset [26], which is one of the most representative benchmarks for aerial image object detection in the remote sensing domain. DOTA-v1 is specifically designed for multi-class object detection tasks in high-resolution remote sensing images. Released by the Aerospace Information Research Institute of the Chinese Academy of Sciences, the dataset includes images from various aerial and satellite imaging platforms, such as Google Earth, GF-2, and JL-1, offering wide coverage, diverse data types, and complex geometric structures.

The dataset contains 2806 high-resolution images, with resolutions ranging from 800 × 800 to 4000 × 4000 pixels and spatial resolutions covering 0.1–1 m/px. It encompasses both dense urban areas with buildings and transportation infrastructure, as well as natural regions including farmland, ports, rivers, and forests, providing high scene diversity and complexity. The dataset includes 15 typical object categories (Fig. 6).

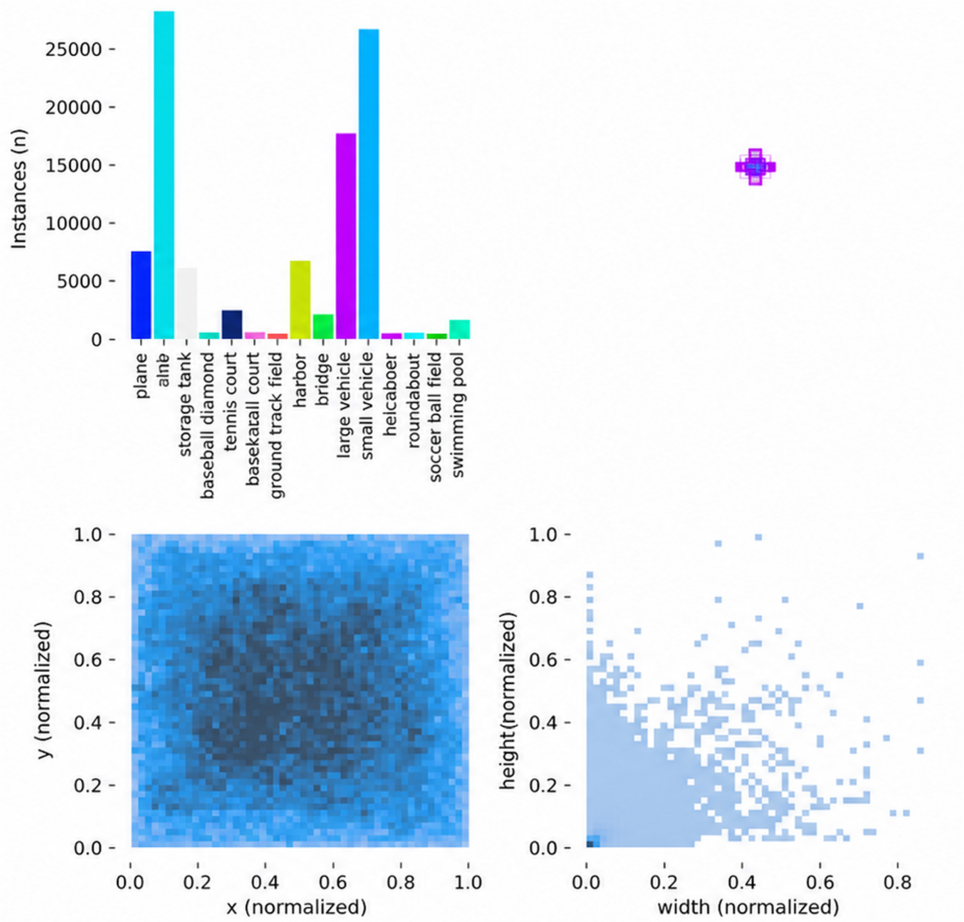


Figure 6: Experimental dataset.

To increase task difficulty and ensure comprehensive algorithm evaluation, the object scale distribution in DOTA-v1 is highly imbalanced. It contains numerous small objects of only a few dozen pixels as well as large objects spanning several hundred pixels. Many areas also feature severe occlusion, dense object

distributions, and background interference, making DOTA-v1 an important benchmark for assessing model robustness and generalization ability.

The DOTA dataset features dual annotations, offering both oriented bounding boxes (OBB) and horizontal bounding boxes (HBB). Nevertheless, as this work is specifically confined to the task of horizontal object detection, we utilize only the HBB annotations for our experiments, leaving the inherent orientation information of the dataset unutilized. The original DOTA-v1 dataset only provides labels for the training and validation sets, while the labels for the test set are not publicly available. Therefore, it is impossible to perform local evaluation strictly following the official split. To enable controlled ablation studies and hyperparameter tuning, we merged the original training and validation sets and re-split them into new training, validation, and test sets with a ratio of 8:1:1. The training set is used for model learning, the validation set for parameter tuning and early stopping, and the test set for final performance evaluation. This strategy ensures consistency and fairness across all experimental evaluations. Consequently, the absolute mAP values reported in this paper may not be directly comparable with those from methods that strictly adhere to the official DOTA-v1 benchmark split. To ensure a fair comparison, all baseline models were retrained under the same 8:1:1 split ratio and experimental settings as our proposed method. Based on this, the relative performance improvements demonstrated in this study are valid and reliable. The main contribution of this work lies in the architectural improvements and their relative gains over baselines under identical conditions.

4.2 Experimental Environment

The experiments in this study were conducted on the AutoDL platform, leveraging high-performance hardware and a stable software environment. The specific configuration is as follows:

- Hardware: NVIDIA RTX 4090 GPU (24 GB) to accelerate model training, paired with an Intel Core i9-14900KF processor and 60 GB of RAM to ensure efficient and stable training.
- Software: PyTorch 2.1.0 with CUDA 12.1 support as the primary deep learning framework, using Python 3.10 for programming.

In the experimental setup, all models are trained for 300 epochs with an input image size of 640 × 640. We use a batch size of 16 and employ the SGD optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005, all of which follow the default settings of the framework. The default online augmentation methods of YOLO are adopted, including random horizontal flipping (with a probability of 0.5) and HSV color space augmentation, with Mosaic augmentation applied throughout the entire training process. The models are initialized using pre-trained weights from ImageNet classification. The random seed is fixed to 0, and deterministic algorithms are enabled. During validation, the confidence threshold and NMS IoU threshold are set to their default values of 0.001 and 0.6, respectively.

4.3 Evaluation Metrics

Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive, reflecting the reliability of the model's positive predictions. It is calculated as follows:

$$P = \frac{TP}{TP + FP}$$

Recall measures the model's ability to capture all true positive samples in the dataset, i.e., how many of the actual positive instances are correctly identified by the model. It is calculated as:

$$R = \frac{TP}{TP + FN}$$

Mean Average Precision (mAP) is used to comprehensively evaluate the detection accuracy of an object detection model across all classes. It is defined as the mean of the Average Precision (AP) values for each class, generally computed as:

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

here, TP, FP, and FN respectively denote the numbers of correctly detected bounding boxes, false positive boxes, and missed boxes. AP represents the area enclosed by the P-R (Precision-Recall) curve and the coordinate axes. mAP50 (i.e., mean Average Precision at IoU = 0.50) is calculated based on an IoU threshold of 0.50 and is used to measure the model's detection accuracy at this threshold; whereas mAP50–95 is computed over IoU values from 0.50 to 0.95 (with a step size of 0.05), providing a more comprehensive reflection of the model's detection performance under different strictness levels and offering greater reference value for evaluating the model's generalization ability and robustness.

5 Experimental Results

5.1 Comparative and Ablation Experiments

This study builds upon the YOLOv11n baseline by introducing three key components: the Large Separable Kernel Attention (LSKA), the Gold-YOLO Neck structure, and the MultiSEAMHead detection head. Based on these components, we construct two improved models: YOLOv11n-LSKA-GoldYOLO and YOLOv11n-GoldYOLO-MultiSEAMHead. To better understand the interactions between different architectural components, we conduct a series of ablation experiments by progressively integrating LSKA, Gold-YOLO, and MultiSEAMHead into the YOLOv11n baseline. These experiments are not predicated on the assumption that each module independently contributes to performance gains; rather, they are designed to investigate the interaction patterns among different modules within a lightweight detection framework and to comprehensively evaluate the performance of the proposed models on the DOTA-v1 dataset.

Table 1 summarizes the key performance metrics, including mAP50, mAP50–95, and Recall. The ablation results indicate that the effectiveness of individual modules does not exist independently, but rather depends on the synergistic interactions among them. Directly integrating the MultiSEAMHead into the baseline YOLOv11n model YOLOv11n+MultiSEAMHead leads to performance degradation. This is due to feature misalignment. The MultiSEAMHead is designed to jointly model multi-level semantic features and fine-grained features. However, the original YOLOv11n neck adopts a relatively simple Feature Pyramid Network structure that provides only limited cross-scale feature interaction. Consequently, the feature maps received by the MultiSEAMHead from different layers are not sufficiently aligned or complementary. The enhanced detection head requires richer and well-aligned multi-level information, but instead becomes "confused" by suboptimal inputs, resulting in increased false positives and false negatives. This is akin to a high-performance engine running on poor-quality fuel. After introducing the Gold-YOLO neck, which provides a sophisticated gathering and distribution mechanism that spatially aligns and enriches features from multiple scales, a "synergistic effect" emerges. This allows the MultiSEAMHead to fully realize its representational capacity, as evidenced by the significant performance improvement of the YOLOv11n-GoldYOLO-MultiSEAMHead variant. Similarly, integrating LSKA with Gold-YOLO also yields performance improvements over the baseline. LSKA enhances the contextual modeling capability of the backbone via a large receptive field, while the Gold-YOLO neck effectively propagates these enhanced features across different scales. It is worth noting that although YOLOv11n+LSKA has a slightly higher number of parameters than the baseline YOLOv11n, its inference speed improves significantly by about 22.5%. This counterintuitive phenomenon stems from improved computational efficiency rather than an increase in

parameter count. Specifically, LSKA replaces standard convolutions with depthwise separable convolutions, effectively reducing memory access overhead and enhancing GPU parallelism. Meanwhile, the optimized feature flow minimizes redundant processing, thereby improving hardware utilization. As a result, despite a slight increase in the number of parameters, the actual inference latency is reduced.

Table 1: Comparison results.

Model	Recall%	mAP50%	mAP50%–95%	Params	GFLOPs	FPS
YOLOv8n	36.6	39.9	24.2	3.2	8.7	30.8
YOLOv9t	37	41.3	25.5	2.0	7.7	35.0
YOLOv10n	35.8	38.6	23.4	2.3	6.7	40.1
YOLOv11n	40.2	42	25.7	2.6	6.3	42.7
YOLOv11n+LSKA	39	42.9	26	2.9	6.5	52.3
YOLOv11n+goldyolo	40.6	42.2	25.8	5.9	9.5	47.8
YOLOv11n+MultiSEAMHead	37.4	40.4	24.2	4.6	6.1	49.2
YOLOv11n+MultiSEAMHead+goldyolo	40.3	43.8	26.3	8.0	9.2	46.9
YOLOv11n+LSKA+goldyolo	38.8	43.3	26	6.2	9.4	46.5

To provide a finer-grained analysis, we further report class-wise AP on selected representative categories, as shown in Table 2. Compared with YOLOv11n, YOLOv11n-GoldYOLO-MultiSEAMHead improves AP on small vehicle from 53.9% to 56.5%, while the AP on ship and plane changes to 34.6% and 59.8%, respectively. YOLOv11n-LSKA-GoldYOLO achieves comparable results to the baseline, with APs of 53.2%, 35.8%, and 60.4% on small vehicle, ship, and plane, respectively. These observations indicate that the gains brought by the proposed modules are category-dependent rather than uniformly consistent across all categories. These results suggest that the improvements are category-dependent rather than universally consistent across all categories.

Table 2: Class-wise mAP50%.

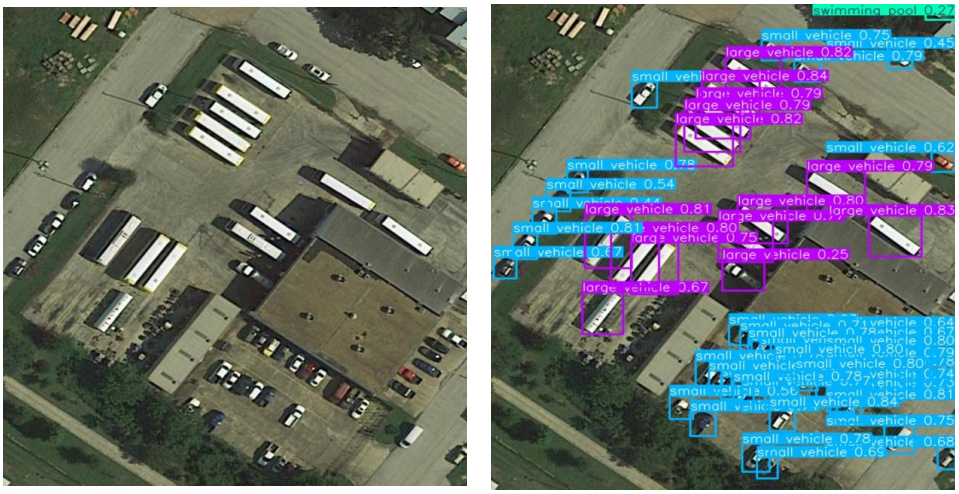
Class	YOLOv11n	YOLOv11n-GoldYOLO-MultiSEAMHead	YOLOv11n-LSKA-GoldYOLO
Small vehicle	53.9%	56.5%	53.2%
Plane	60.1%	59.8%	60.4%
Ship	36.5%	34.6%	35.8%

Considering both detection performance and computational complexity, two model variants are ultimately retained in this study. Compared to the YOLOv11n baseline, YOLOv11n-LSKA-GoldYOLO achieves improvements of 1.3% in mAP50 and 0.3% in mAP50–95, while YOLOv11n-GoldYOLO-MultiSEAMHead achieves improvements of 1.8% and 0.6%, respectively. Although the additional modules increase the number of parameters, the overall model size and computational cost remain low, indicating that the proposed method maintains its lightweight advantage.

When real-time inference and global context awareness are the main concerns and computational resources are limited, YOLOv11n-LSKA-GoldYOLO is recommended. This model has a moderate number of parameters (6.2M, 2.38 times that of the baseline), a stable FPS of 46.5, a 9% speed improvement over the baseline, and a 1.3% mAP50 gain, making it highly suitable for deployment on edge devices such as onboard drone computers. Its LSKA module enhances background suppression capability, which is particularly effective in sparse small-target scenarios like suburban vehicle detection. If higher accuracy

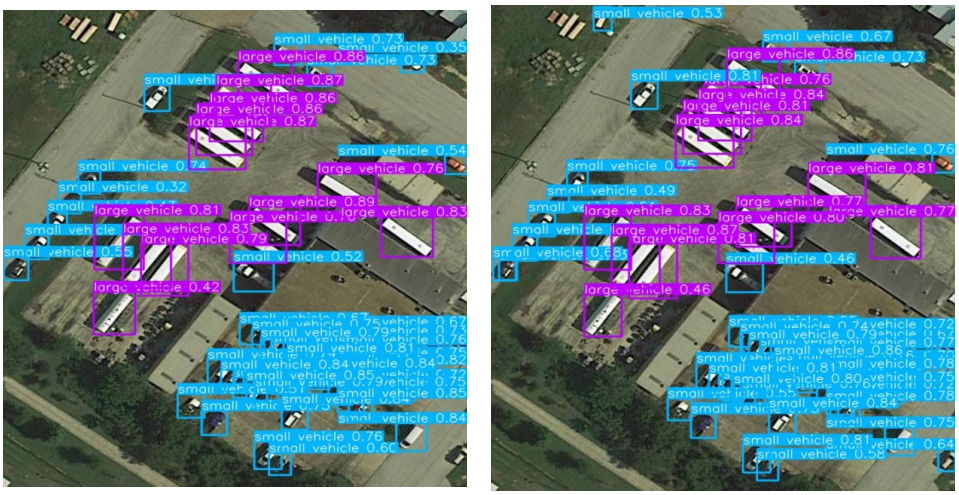
is desired and a slightly higher computational cost is acceptable, YOLOv11n-GoldYOLO-MultiSEAMHead is recommended. This model achieves a 1.8% mAP50 improvement (reaching 43.8%) and an FPS of 46.9, performing excellently in dense, occluded, or highly variable scenes such as port management and urban traffic monitoring. However, its parameter count increases to 8.0M (3.08 times that of the baseline), making it more suitable for deployment on server-grade GPUs or cloud platforms, or for offline aerial image analysis tasks where real-time requirements are not stringent.

Fig. 7 shows original images from the DOTA-v1 dataset and detection results of the YOLOv11n model. Fig. 8a presents the detection results of the YOLOv11n-LSKA-GoldYOLO model on the validation set, while Fig. 8b shows the detection results of the YOLOv11n-GoldYOLO-MultiSEAMHead model on the validation set.



(**a**) (**b**)

Figure 7: Qualitative detection results on the DOTA-v1 dataset. (**a**) Original image. (**b**) Detection results of the baseline YOLOv11n model.



(**a**) (**b**)

Figure 8: Qualitative detection results of the proposed models. (**a**) YOLOv11n-LSKA-GoldYOLO. (**b**) YOLOv11n-GoldYOLO-MultiSEAMHead.

Both proposed models outperform the baseline YOLOv11n in detection performance, yet they emphasize different architectural designs and practical advantages. YOLOv11n-LSKA-GoldYOLO enhances the backbone network by introducing a large-receptive-field attention mechanism, making it particularly suitable for tasks where global contextual information is critical, such as scenarios involving small, sparsely distributed targets or objects embedded in complex backgrounds. Since it introduces relatively fewer modifications to the detection head, it incurs lower computational overhead and is therefore more appropriate for deployment environments with strict real-time requirements or limited computational resources. In contrast, YOLOv11n-GoldYOLO-MultiSEAMHead focuses more on strengthening feature interaction during the detection stage. By enhancing multi-level semantic fusion and spatial–channel attention modulation, this model demonstrates greater robustness in scenarios with dense targets, occlusion, or significant scale variations. Therefore, it is better suited for high-precision detection tasks in urban or port environments where object overlap and background clutter are common. By providing two targeted variants, we enable flexible selection of a more appropriate model according to specific application requirements.

5.2 Robustness Analysis under Noise Perturbations

To evaluate the robustness of the proposed model in real-world complex scenarios, we conducted additional experiments by adding increasing levels of Gaussian noise (GN10–GN50) to the test set images, simulating realistic interference conditions that cannot be fully captured by the standard DOTA-v1 metrics. Table 3 shows the detection performance (mAP50) of different models under various noise levels.

Table 3: Robustness comparison under Gaussian noise perturbations (mAP50%).

Model	Clean	GN10	GN20	GN30	GN40	GN50	Avg-GN	Retention (%)
YOLOv11n	42%	39.8%	28.5%	18.4%	10.6%	5.6%	20.6%	49.0%
YOLOv11n+MultiSEAM Head+goldyolo	43.8%	41.1%	30.2%	17.6%	10.0%	6.8%	21.1%	48.3%
YOLOv11n+LSKA+goldyolo	43.3%	39.8%	30.2%	18.7%	11.8%	8.1%	21.7%	50.2%

Compared with the baseline model YOLOv11n, both improved models achieved higher average performance under noisy conditions. Among them, YOLOv11n-LSKA-GoldYOLO obtained the highest Avg-GN (21.7%), which is 1.1% higher than that of the baseline, indicating its stronger robustness to noise interference. Specifically, under severe noise (GN40–GN50), this model significantly outperformed YOLOv11n in detection accuracy, demonstrating that the large receptive field modeling introduced by LSKA effectively enhances global contextual awareness and suppresses noise interference.

In contrast, YOLOv11n-GoldYOLO-MultiSEAMHead performed better under moderate noise (GN10–GN20), but its robustness slightly decreased at higher noise levels. This suggests that although MultiSEAMHead enhances feature representation, it becomes more sensitive to feature quality degradation when the input noise is severe.

The retention rate reflects the model's ability to maintain performance under noisy environments. Compared with the baseline, YOLOv11n-LSKA-GoldYOLO achieved the highest retention rate (50.2%), further validating its superior robustness.

In summary, the proposed improvements not only increase detection accuracy but also significantly enhance model robustness under complex conditions, making the model more suitable for practical remote sensing applications.

6 Conclusion

This study addresses the challenges of large-scale variation, dense distribution, and complex backgrounds in aerial remote sensing image object detection by proposing improved lightweight object detection models based on YOLOv11n, namely YOLOv11n-LSKA-GoldYOLO and YOLOv11n-GoldYOLO-MultiSEAMHead. By introducing the Large Separable Kernel Attention (LSKA) mechanism into the backbone network, the models significantly enhance global receptive field and contextual information modeling capabilities. Incorporating the Gold-YOLO Neck during feature fusion enables more efficient and stable multi-scale feature interactions, and the integration of the MultiSEAMHead module achieves deep cross-layer feature fusion with stronger spatial–channel joint modeling capability.

Experimental results on the DOTA-v1 dataset demonstrate that the proposed improve the mAP-based detection accuracy over the baseline, although Recall does not consistently increase across all variants, indicating that appropriate module combinations can improve the overall detection performance of YOLOv11n on DOTA-v1, while the class-wise gains are not uniform across all categories. In practical applications, YOLOv11n-LSKA-GoldYOLO is more suitable for lightweight deployment scenarios requiring enhanced global contextual perception, whereas YOLOv11n-GoldYOLO-MultiSEAMHead is preferable for complex and densely distributed scenes demanding higher detection robustness. The two models provide flexible and extensible solutions for different remote sensing application requirements.

Although the methods proposed in this paper achieve improvements, several limitations remain. First, performance gains are category-dependent: LSKA improves detection of context-rich objects (e.g., airplanes) but may slightly degrade detection of tiny, texture-poor objects (e.g., small vehicles); the MultiSEAMHead variant also shows limited improvement for ships. The parameter count of the new model increases, which may hinder its deployment on extremely resource-constrained devices (e.g., micro-UAVs). Recall does not improve—the LSKA variant even shows slightly lower recall than the baseline, while MultiSEAMHead only partially alleviates this issue. All experiments are conducted on the DOTAv1 dataset; its generalizability to other remote sensing datasets remains to be verified. Future work should focus on addressing these issues.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Shuaiyu Zhu and Ji Li; methodology, Shuaiyu Zhu and Ji Li; software, Shuaiyu Zhu and Ji Li; validation, Shuaiyu Zhu and Ji Li; formal analysis, Shuaiyu Zhu and Ji Li; investigation, Shuaiyu Zhu and Ji Li; resources, Shuaiyu Zhu and Ji Li; data curation, Shuaiyu Zhu and Ji Li; writing—original draft preparation, Shuaiyu Zhu; writing—review and editing, Shuaiyu Zhu; visualization, Shuaiyu Zhu and Ji Li; supervision, Sergey Ablameyko; project administration, Sergey Ablameyko; funding acquisition, Shuaiyu Zhu and Sergey Ablameyko. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data underlying this article will be shared on reasonable request to the corresponding author.

Ethics Approval: Not applicable.

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

References

1. Zhao Y, Rao Y, Dong S, Zhang J. Survey on deep learning object detection. J Image Graph. 2020;25(4):629–54. doi:10.11834/jig.190307.

2. Zhiyuan L, Hao W, Ma G, Weichen Y, Ablameyko S. Effective small object detection in remote sensing images based on improved YOLOv8 network. Nonlinear Phenom Complex Syst. 2024;27(3):278–91. doi:10.1109/access.2020.3021895.
3. Chen XF, Li M, Zhao JY, Lyu YL, He YJ. Remote sensing small object detection algorithm based on dual-layer attention mechanism. J Rocket Force Univ Eng. 2025;39(1):60–6. (In Chinese). doi:10.1109/ccdc65474.2025.11090382.
4. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. Adv Neural Inf Process Syst. 2021;34:15908–19.
5. Jiang P, Ergu D, Liu F, Cai Y, Ma B. A review of yolo algorithm developments. Procedia Comput Sci. 2022;199(11):1066–73. doi:10.1016/j.procs.2022.01.135.
6. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Proceedings of the Computer vision—ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
7. Xu GD, Mao GJ. Aerial image object detection of UAV based on multi-level feature fusion. J Front Comput Sci Technol. 2023;17(3):635–45. (In Chinese). doi:10.1007/s11760-026-05291-9.
8. Du Z, Hu Z, Zhao G, Jin Y, Ma H. Cross-layer feature pyramid transformer for small object detection in aerial images. IEEE Trans Geosci Remote Sens. 2025;63:5625714. doi:10.1109/TGRS.2025.3572706.
9. Zhou S, Zhou H, Qian L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. Sci Rep. 2025;15(1):9255. doi:10.1038/s41598-025-92344-7.
10. Song J, Miao L, Ming Q, Zhou Z, Dong Y. Fine-grained object detection in remote sensing images via adaptive label assignment and refined-balanced feature pyramid network. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:71–82. doi:10.1109/JSTARS.2022.3224558.
11. Yuan H, Zhang B, Wang Y, Qiang Q. From structural degradation to semantic misalignment: a unified frequency-aware compensation framework for remote sensing object detection. Remote Sens. 2026;18(5):777. doi:10.3390/rs18050777.
12. Li H, Qu H. VMC-Net: multi-scale feature aggregation and distribution with contextual attention guided fusion for aerial object detection. Complex Intell Syst. 2025;11(8):350. doi:10.1007/s40747-025-01888-8.
13. Fu T, Dong H, Yang B, Deng B. DE-DFNet: edge enhanced diversity feature fusion guided by differences in remote sensing imagery tiny object detection. Image Vis Comput. 2025;161(9):105627. doi:10.1016/j.imavis.2025.105627.
14. Liu L, Li J. MCRS-YOLO: multi-aggregation cross-scale feature fusion object detector for remote sensing images. Remote Sens. 2025;17(13):2204. doi:10.3390/rs17132204.
15. Lau KW, Po LM, Rehman YAU. Large separable kernel attention: rethinking the large kernel attention design in CNN. Expert Syst Appl. 2024;236(9):121352. doi:10.1016/j.eswa.2023.121352.
16. Wu Z, Zhen H, Zhang X, Bai X, Li X. SEMA-YOLO: lightweight small object detection in remote sensing image via shallow-layer enhancement and multi-scale adaptation. Remote Sens. 2025;17(11):1917. doi:10.3390/rs17111917.
17. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 (Version 8.0.0). 2023 [cited 2026 Jan 1]. Available from: https://github.com/ultralytics/ultralytics.
18. Rao SN. YOLOv11 architecture explained: next-level object detection with enhanced speed and accuracy. Medium. 2024 [cited 2026 Jan 1]. Available from: https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speedand-accuracy-2dbe2d376f71.
19. Guo J, Han K, He W, Liu C, Nie Y, Wang C, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism. In: Proceedings of the Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 51094–112. doi:10.52202/075280-2224.
20. Wan Q, Huang Z, Lu J, Yu G, Zhang L. SeaFormer++: squeeze-enhanced axial transformer for mobile visual recognition. Int J Comput Vis. 2025;133(6):3645–66. doi:10.1007/s11263-025-02345-2.
21. Huang Z, Ben Y, Luo G, Cheng P, Yu G, Fu B. Shuffle transformer: rethinking spatial shuffle for vision transformer. arXiv:2106.03650. 2021.
22. Xiao Y, Xu T, Yu X, Fang Y, Li J. A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection. IEEE Trans Geosci Remote Sens. 2024;62:4708011. doi:10.1109/TGRS.2024.3457155.

23. Huang F, Liu H, Chen L, Shen Y, Yu M. Feature enhanced cascading attention network for lightweight image super-resolution. Sci Rep. 2025;15(1):2051. doi:10.1038/s41598-025-85548-4.
24. Qu J, Tang Z, Zhang L, Zhang Y, Zhang Z. Remote sensing small object detection network based on attention mechanism and multi-scale feature fusion. Remote Sens. 2023;15(11):2728. doi:10.3390/rs15112728.
25. Tang H, Jiang Y. An improved YOLOv8n algorithm for object detection with CARAFE, MultiSEAMHead, and TripleAttention mechanisms. In: Proceedings of the 2024 7th International Conference on Computer Information Science and Application Technology (CISAT); 2024 Jul 12–14; Hangzhou, China. p. 119–22. doi:10.1109/CISAT62382.2024.10695221.
26. Xia GS, Bai X, Ding J, Zhu Z, Belongie S, Luo J, et al. DOTA: a large-scale dataset for object detection in aerial images. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 3974–83. doi:10.1109/CVPR.2018.00418.