An Intelligent Agent-Based Multi-Scale Target Detection System for Remote Sensing Images Using YOLOv8

Jifeng Ding, Jiayuan Lin, Kechao Zhang and Xuan Wang

Abstract—**Remote sensing technology has been widely adopted in resource exploration, environmental monitoring, and other domains, making object detection in remote sensing imagery an increasingly critical research area. However, significant challenges persist due to variable object scales and complex backgrounds in these images. This thesis introduces GMP-YOLO, an enhanced YOLOv8-based model specifically designed to improve detection performance for small and multi-scale objects in remote sensing imagery. The proposed methodology incorporates several key innovations: (1) improving the CIoU loss function of the original YOLOv8 to PIoUv2 loss function, which optimizes anchor box regression and expedites model convergence; (2) designing a Multi-Scale High-Resolution Feature Fusion Network (MSHR-Net) containing a Multi-scale Detection Block (MDB) that enhances small object recognition and manages complex backgrounds; and (3) introducing a Global Attention Mechanism (GAM) to strengthen feature extraction capabilities. Quantitative analyses demonstrate that GMP-YOLO achieves a 7.0% improvement in average detection accuracy on the DOTA dataset compared to the baseline model. The proposed architecture also exhibits exceptional performance on the VisDrone dataset, particularly in complex environmental settings. Furthermore, an intelligent agent system developed with this model demonstrates superior real-time analysis capabilities for urban remote sensing imagery, providing valuable support for urban planning and management applications. These findings confirm that GMP-YOLO not only delivers enhanced efficiency and robustness for remote sensing object detection tasks but also presents a promising solution for intelligent urban remote sensing systems.**

Index Terms—**Remote Sensing Image, Object Detection, Multi-Scale Detection, Agent, YOLOv8**

I. Introduction

Remote sensing technology has emerged as a valuable tool for large-scale, all-weather, and high-precision observations, particularly following advancements in high-resolution imagery acquisition. The resulting high-resolution remote sensing datasets, with their rich spatial, spectral, and temporal information, have become essential for resource exploration, urban planning, environmental monitoring, and national security applications [1]. However, the rapid extraction of accurate target information from these increasingly complex remote sensing images remains a critical challenge in the field.

Object detection research in remote sensing has evolved alongside mainstream computer vision, transitioning from rule-based approaches to deep learning frameworks. Despite this progress, two significant issues persist: small object detection and multi-scale object detection. Small objects occupy minimal pixel areas, leading to information loss during feature extraction [2]. Their detection is further complicated by complex backgrounds and variable scale distributions. Conversely, multi-scale detection must address the extensive scale variation in remote sensing imagery, from small buildings to large terrain features, requiring simultaneous capture of fine details and broader structural contexts [3].

Deep learning has significantly advanced remote sensing object detection through end-to-end feature learning. Two-stage methods like Fast R-CNN employ Region Proposal Networks (RPNs) to generate candidate regions before classification, enhancing accuracy at the cost of processing speed [4]. To address this limitation, single-stage detectors emerged. The Single Shot MultiBox Detector (SSD) performs detection across multi-scale feature maps, effectively balancing accuracy and efficiency [5]. The YOLO (You Only Look Once) architecture reformulates detection as a regression problem, substantially improving computational efficiency [6]. Recent variants such as YOLOv4 and YOLOv5 have further optimized network architectures and training strategies for diverse hardware environments [7].

Despite their success with natural images, these models face considerable difficulties when applied to remote sensing data. The unique characteristics of remote sensing imagery—particularly the complex distribution of small and multi-scale objects—often exceed conventional detection capabilities [8]. Consequently, researchers have focused on adapting existing frameworks for remote sensing applications. Etten et al. pioneered YOLO implementation in this domain [9], while Huang et al. enhanced YOLOv4 with dilated convolution modules to improve small object detection in complex backgrounds [10]. Du and Liang modified YOLOv5 by incorporating specialized detection

Manuscript received December 26, 2024; revised April 26, 2025.

Jifeng Ding is an associate professor at the School of Information and Communication Engineering, Dalian Minzu University, Dalian, China. (Corresponding author, phone: +86-186-4095-2660; e-mail: djf@dlnu.edu.cn).

Jiayuan Lin is a postgraduate student at the School of Information and Communication Engineering, Dalian Minzu University, Dalian, China. (email: 202211051055@stu.dlnu.edu.cn).

Kechao Zhang is a postgraduate student at the School of Information and Communication Engineering, Dalian Minzu University, Dalian, China. (e-mail: 202211055092@stu.dlnu.edu.cn).

Xuan Wang is an undergraduate student at the School of Information and Communication Engineering, Dalian Minzu University, Dalian, China. (e-mail: 2022136422@stu.dlnu.edu.cn).

heads and optimizing feature fusion to reduce information loss during down-sampling [11]. While these approaches have advanced the field, they typically target specific feature optimizations rather than comprehensively addressing the diverse scales and complex backgrounds inherent in remote sensing imagery. Balancing detection performance across variable target scales remains a significant hurdle in multi-scale detection scenarios.

Recent studies have attempted to address this challenge through innovative architectural modifications. Zhang et al. integrated EfficientNetV2 and C3Ghost networks into YOLOv5s with a Shuffle attention mechanism, enhancing multi-scale detection while reducing computational complexity [12]. Wei et al. proposed MTD-YOLOv5, featuring a multi-scale perceptual hybrid pooling module that combines horizontal and vertical receptive fields to capture target information more effectively in complex environments [13]. Despite these advances, achieving balanced performance between small and large object detection continues to present significant obstacles.

Simultaneously, the application landscape for remote sensing imagery is evolving toward greater intelligence, particularly in smart city management contexts where efficient information extraction from extensive datasets is crucial for urban decision-making.

To address these challenges, this paper proposes an improved model, GMP-YOLO, based on YOLOv8. GMP-YOLO introduces innovative enhancements in model architecture, feature extraction, and target detection modules, aiming to achieve a balance between small target detection and multi-scale target detection in remote sensing imagery. The specific improvements are as follows:

1) Adoption of the upgraded Powerful-IoU (PIoU v2) loss function [14], which enhances anchor box regression accuracy and accelerates model convergence by incorporating a non-monotonic focusing strategy.
2) Design of a Multi-scale Detection Block (MDB), based on the Receptive Field Block (RFB) framework [15], augmented with the SE attention mechanism. This improves the model's feature extraction capabilities across different target scales, effectively boosting multi-scale target detection performance.
3) Development of a multi-scale high-resolution feature fusion network (MSHR-Net), combined with the path-aggregation feature pyramid network (PA-FPN). This includes the addition of a new detection scale for tiny targets and optimizations to preserve feature information during the downsampling process, significantly improving small target detection.
4) Incorporation of the GAM module [16] into the backbone network to enhance multi-dimensional feature capture, further improving the model's detection performance in complex backgrounds.

II. Materials and Method

YOLOv8, developed by Ultralytics, represents a significant advancement in object detection technology. Despite the emergence of newer frameworks in recent years, YOLOv8 maintains its position as a crucial solution for diverse real-world applications due to its exceptional balance between accuracy and computational efficiency. Its architecture, illustrated in Fig. 1, comprises three principal components: Backbone, Neck, and Head. This modular structure provides an excellent foundation for the model's adaptability and extensibility while ensuring operational stability and performance in complex detection environments.

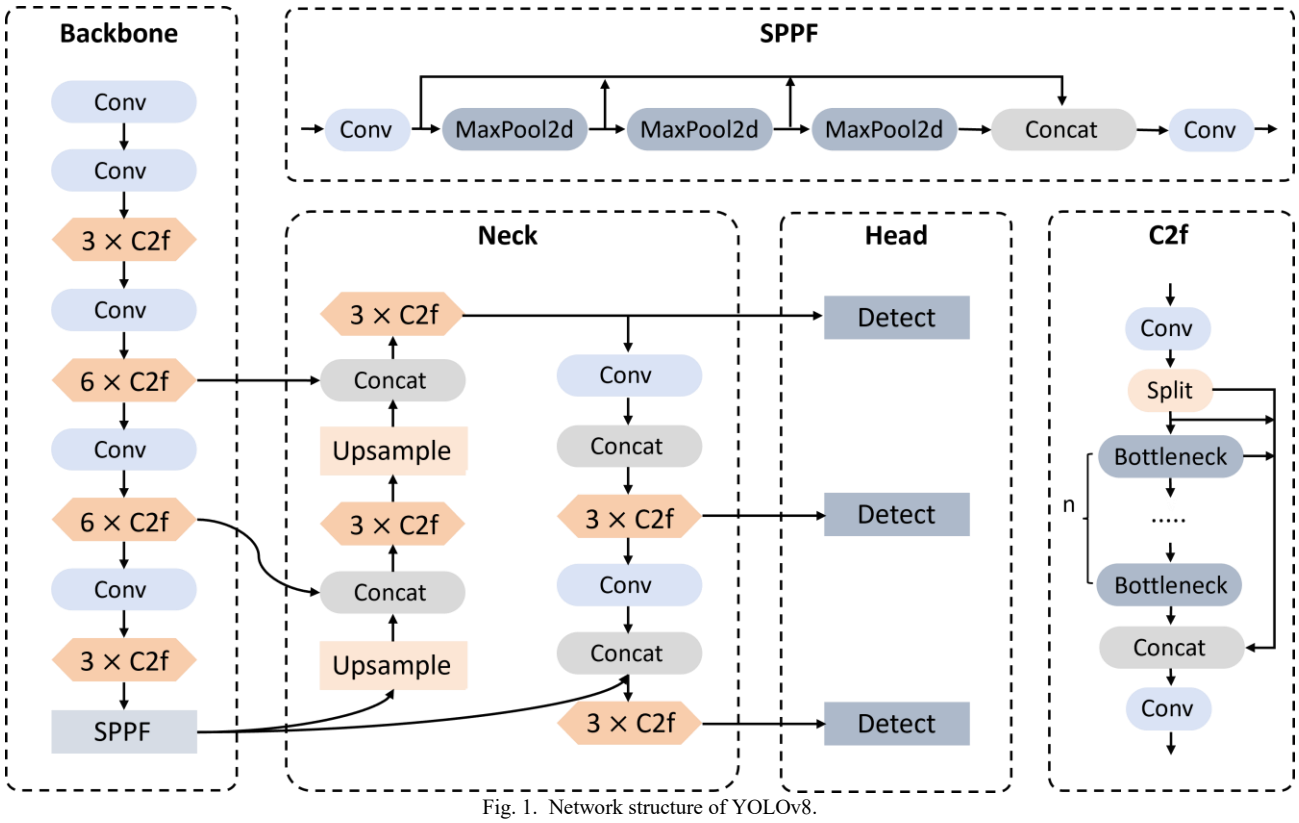


Fig. 1. Network structure of YOLOv8.

The backbone network of YOLOv8 employs an enhanced version of CSPDarknet-53 as its foundation, integrating the C2f module [17] to supersede the conventional CSP module. The C2f module augments detection capabilities for small targets by effectively merging high-level semantic features with low-level spatial information. At the terminal phase of the backbone, the SPPF module [18] is utilized to generate fixed-length feature vectors through multi-scale pooling operations. This component improves upon the traditional SPP architecture by optimizing its configuration to decrease computational demands while simultaneously enhancing processing efficiency.

The neck structure implements the Path Aggregation Feature Pyramid Network (PA-FPN) architecture, which combines top-down and bottom-up feature fusion paths to effectively integrate cross-level information. To further boost computational performance and minimize complexity, YOLOv8 refines the traditional PAN architecture by eliminating the convolutional layer following the upsampling operation. This modification results in a streamlined design, substantially improving support for real-time detection applications while preserving high performance metrics.

In the detection head, YOLOv8 adopts a decoupled head strategy [19], separating the classification and bounding box regression tasks into two independent branches. The classification branch utilizes a binary cross-entropy loss function to ensure accurate target categorization. Concurrently, the bounding box regression branch incorporates the distribution focus loss and the CIoU loss function to enhance the accuracy and robustness of bounding box predictions. This configuration improves the model's detection capabilities in challenging scenarios while optimizing overall inference efficiency.

III. Improved Strategy

Despite YOLOv8's robust performance in general object detection tasks, its standard framework exhibits limitations when applied to multi-scale targets in remote sensing imagery, particularly regarding feature extraction precision and computational efficiency. To address these constraints, this study proposes GMP-YOLO, an enhanced model based on YOLOv8 with several key architectural improvements.

To enhance the effectiveness and precision of bounding box regression, GMP-YOLO implements the PIoUv2 loss function. By introducing a non-monotonic focusing strategy, PIoUv2 guides anchor boxes along optimized regression paths, accelerating model convergence while improving bounding box localization accuracy. Based on the Receptive Field Block framework, the Multi-scale Detection Block (MDB) enhances detection performance across targets of various dimensions. By integrating MDB with the PA-FPN architecture in YOLOv8, we developed the Multi-Scale High-Resolution Feature Fusion Network (MSHR-Net). This network enhances multi-scale target detection capabilities by introducing a dedicated detection scale for tiny targets, enabling precise detection across four distinct scales and significantly improving the model's capacity to process complex multi-scale targets in remote sensing imagery. Furthermore, to mitigate potential information degradation in deep network layers, GMP-YOLO incorporates the Global Attention Mechanism (GAM) into the backbone network. GAM substantially enhances the model's ability to capture and preserve critical details, improving detection performance in complex scenarios.

The comprehensive architecture of the proposed GMP-YOLO model is illustrated in Fig. 2.

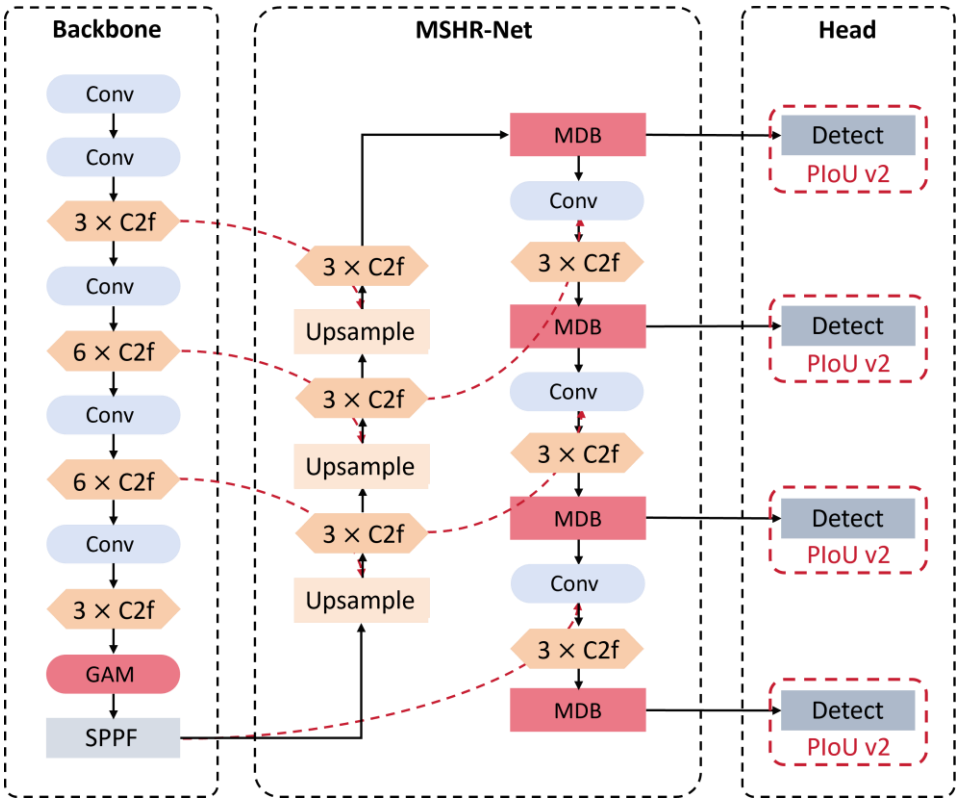


Fig. 2. Network structure of GMP-YOLO.

A. Loss Function Improvement

To develop a remote sensing image target detection model that balances detection accuracy and speed, this study optimizes the loss function for Bounding Box Regression (BBR). BBR plays a critical role in target detection, with the design of its loss function directly influencing model performance. In YOLOv8, DFL loss and CIoU loss [20] are employed for BBR. The formula for CIoU loss is as follows:

$$\mathrm{L}_{CIoU} = 1 - \mathrm{L}_{IoU} + \frac{\rho^2\left(\mathrm{B}_{gt}, \mathrm{B}_{\mathrm{prd}}\right)}{C^2} + \alpha V \tag{1}$$

$$\alpha = \frac{V}{1 - IoU + V}, V = \frac{4}{\pi^2}\left(arctan\frac{w_{gt}}{h_{gt}} - arctan\frac{w_{prd}}{h_{prd}}\right)^2 \tag{2}$$

Here, IoU represents the intersection-over-union between the ground truth bounding box B_{gt} and the predicted bounding box $\mathrm{B}_{\mathrm{prd}}$ · $\rho^2\left(\mathrm{B}_{gt}, \mathrm{B}_{\mathrm{prd}}\right)$ denotes the Euclidean distance between the centroids of the two bounding boxes.

Compared to IoU, GIoU, and DIoU, CIoU incorporates center point distance and aspect ratio differences for a more comprehensive evaluation. However, the use of inverse trigonometric functions increases computational complexity. Moreover, CIoU's penalty mechanism does not directly reflect shape differences in bounding boxes, potentially leading to poor convergence and unreasonable anchor box expansion.

To address these limitations, this study introduces the Powerful-IoU (PIoU) loss function, which incorporates a target size adaptation penalty factor and a gradient adjustment function to optimize the anchor box convergence path. The penalty factor P is defined as:

$$P = \left(\frac{dw_1}{w_{gt}} + \frac{dw_2}{w_{gt}} + \frac{dh_1}{h_{gt}} + \frac{dh_2}{h_{gt}}\right)/4 \tag{3}$$

Here, dw_1, dw_2, dh_1, and dh_2 are the absolute distances between the predicted bounding box and the edges of the ground truth bounding box, while w_{gt} and h_{gt} represent the width and height of the ground truth bounding box. Unlike other methods, p depends solely on the bounding box size and is independent of the enclosing rectangle size.

The gradient adjustment function $f\left(x\right)$ is defined as:

$$f\left(x\right) = 1 - e^{-x^2} \tag{4}$$

This function reduces gradient updates for both high- and low-quality anchor boxes to avoid over-updating, while assigning larger gradient values to medium-quality anchor boxes to accelerate their improvement. The combined PIoU loss is given as:

$$\mathrm{L}_{PIoU} = 1 - PIoU = \mathrm{L}_{IoU} + 1 - e^{-P^2}, 0 \le \mathrm{L}_{PIoU} \le 2 \tag{5}$$

To further address sample imbalance, PIoU v2 introduces a non-monotonic focusing function with a single hyperparameter q. This function emphasizes medium-quality anchor boxes, improving convergence. The focusing function is defined as:

$$q = e^{-P}, q \in \left(0, 1\right] \tag{6}$$

$$u\left(x\right) = 3x \cdot e^{-x^2} \tag{7}$$

The PIoU v2 loss equation is:

$$\mathrm{L}_{PIoU\_v2} = u\left(\lambda q\right) \cdot \mathrm{L}_{PIoU} = 3 \cdot \left(\lambda q\right) \cdot e^{-(\lambda q)^2} \cdot \mathrm{L}_{PIoU} \tag{8}$$

Here, $u\left(\lambda q\right)$ represents the focusing function, where q is the key hyperparameter that adjusts the behavior of the focusing function. The value of q is determined based on the penalty factor P, which evaluates the quality of the anchor box.

Compared to CIoU, PIoU v2 simplifies parameter tuning while enhancing the focus on medium-quality anchor boxes through its non-monotonic focusing strategy. This design effectively improves the convergence efficiency and detection performance of the model. As a result, PIoU v2 is selected as the core loss function for BBR in this study.

B. Improvements in Network Structure

Effective recognition of small targets and adaptation to multi-scale objects remain critical challenges in remote sensing image target detection. To address these limitations, this study enhances the PA-FPN architecture of YOLOv8 and introduces an innovative Multi-scale Detection Block (MDB) and a Multi-scale High-Resolution Feature Fusion Network (MSHR-Net). These improvements significantly enhance multi-scale target detection performance.

The Receptive Field Block (RFB) module simulates the sensory field mechanism of the human visual system through multi-branch pooling and dilated convolution, achieving robust performance in lightweight neural networks, as illustrated in Fig. 3.

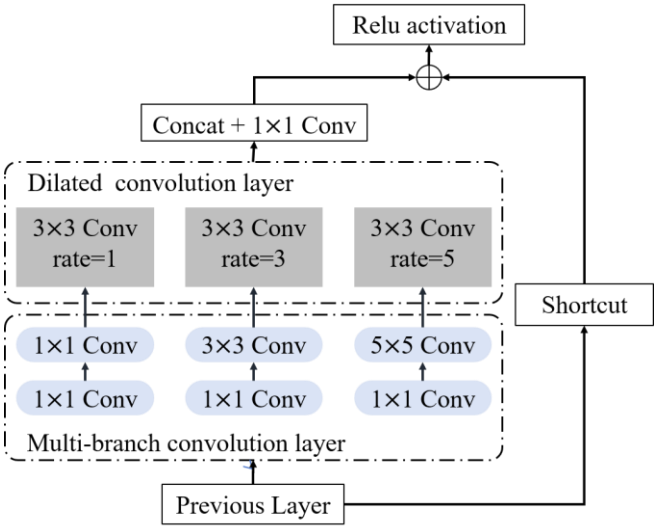


Fig. 3. Network structure of RFB.

However, its efficacy diminishes in remote sensing image scenarios, which typically feature high resolution, complex backgrounds, and densely distributed small targets. Consequently, the RFB module struggles to comprehensively capture detailed multi-scale target information. To overcome these limitations, we propose the Multi-scale Detection Block (MDB), specifically optimized for the unique characteristics of remote sensing imagery.

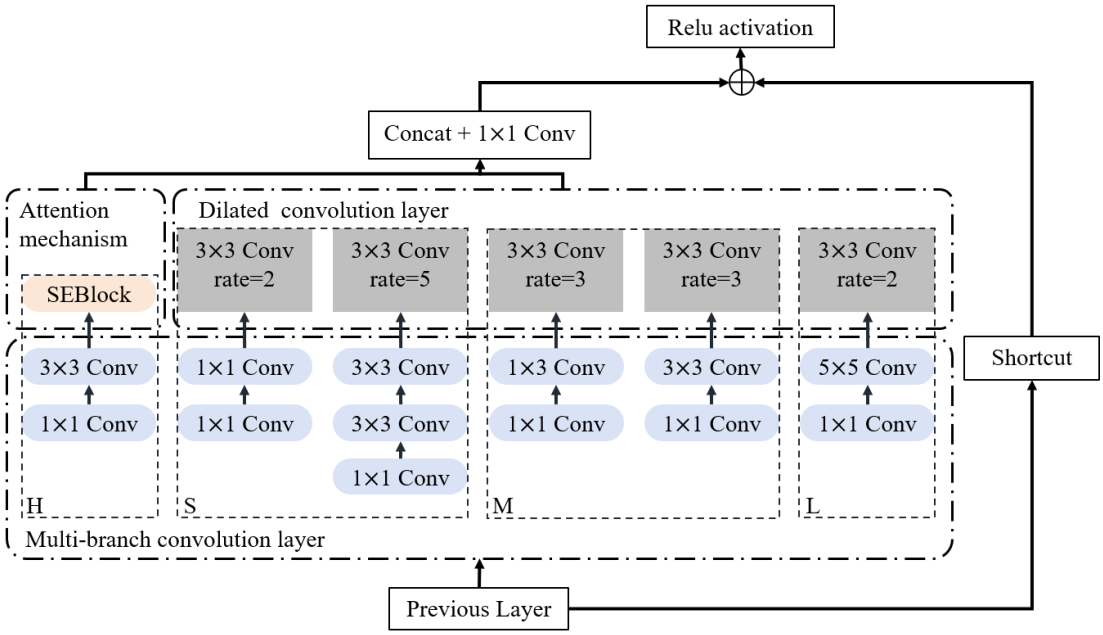


Fig. 4. Network structure of MDB.

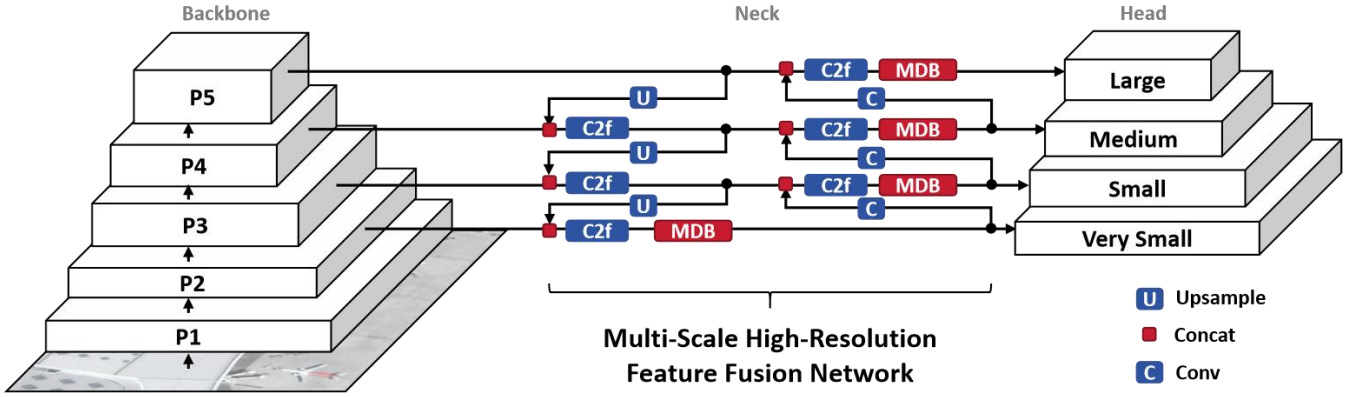


Fig. 5. Architecture diagram of MSHR-Net.

As depicted in Fig. 4, the MDB module is engineered to address the diversity and dimensional variation of targets in remote sensing images, with particular emphasis on small target detection. To process the rich details in high-resolution remote sensing data, MDB incorporates a specialized high-resolution processing branch. This branch combines 1×1 and 3×3 convolutions, where the former reduces dimensionality while the latter extracts spatial features. Additionally, this branch integrates the Squeeze-and-Excitation Block (SE Block) [21], which employs an attention mechanism to amplify the significance of key features, thereby enhancing feature representation quality.

For small target detection, MDB incorporates two specialized pathways. The first pathway combines two standard convolutional layers with a dilated convolution (expansion rate = 2), enhancing feature capture capability while maintaining computational efficiency. The second pathway employs a deeper network with a 3×3 convolutional kernel to capture complex contextual information, further improving the representation of small target features.

For medium-sized target detection, MDB utilizes asymmetric 1×3 and standard 3×3 convolutional kernels. The asymmetric kernel enhances adaptability to targets with varying aspect ratios, while the standard kernel improves detection of targets with regular shapes. For large targets, MDB combines a larger 5×5 convolutional kernel with appropriately configured dilated convolution, allowing for an expanded receptive field and the capture of broader contextual information. This integration of multi-sized convolutional kernels with flexible dilation rates enables MDB to efficiently process small, medium, and large targets, excelling in multi-scale detection tasks.

Building upon the MDB module, this study further designs the Multi-scale High-Resolution Feature Fusion Network (MSHR-Net) to enhance small target detection and comprehensively improve multi-scale detection capabilities. As illustrated in Fig. 5, MSHR-Net refines the Neck structure of YOLOv8 by enabling feature maps to connect with the P3 feature map of the backbone through two upsampling operations and a Concat operation. To further strengthen small target detection in remote sensing images, an additional upsampling step is incorporated in MSHR-Net. This step adjusts the resolution of the feature map to match that of the P2 feature map for subsequent feature concatenation.

The P2 feature map, derived from a shallower network layer, possesses higher spatial resolution and preserves more detailed information, making it particularly advantageous for

detecting small-sized targets. The fused features are processed through the C2f module and subsequently refined by the MDB module. The MDB-processed features are divided into two pathways. The first pathway directly delivers features to the Head, establishing a new detection scale dedicated to tiny target detection. The second pathway further processes these features through additional convolutional operations. This architecture effectively combines detailed spatial information from shallow layers with rich semantic features from deeper layers, significantly enhancing overall detection performance.

To further strengthen MSHR-Net's capability in handling multi-scale targets, the MDB module is integrated into the three original detection scales in the Head. These modifications enable MSHR-Net to efficiently address the challenges of detecting tiny targets while maintaining robust feature capture across multiple scales. Through these architectural innovations, MSHR-Net achieves superior performance in remote sensing image target detection tasks.

C. *Effective Attention Mechanism*

While the Multi-scale High-Resolution Feature Fusion Network (MSHR-Net) significantly improves the model's ability to detect targets in remote sensing images, the increase in model depth can lead to the loss of critical information. To address this issue, we introduce a Global Attention Mechanism (GAM) into the backbone network. GAM enhances the overall performance of the network by reducing information loss and promoting global feature interaction. This attention mechanism builds upon the sequential channel-space attention arrangement from CBAM, with submodules innovatively designed to efficiently capture and enhance global interaction characteristics. This results in improved capability to process complex information. The specific structure of GAM is shown in Fig. 6.

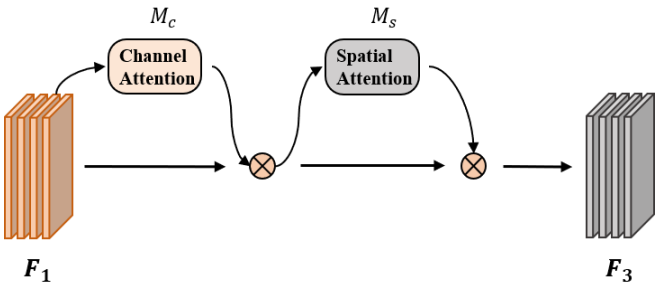


Fig. 6. Structure of GAM.

As illustrated in Fig. 6, GAM consists of two components: Channel Attention and Spatial Attention. The feature map F_1 extracted by the C2f module is passed into GAM for processing, producing an intermediate state F_2 and an output F_3, which can be expressed as:

$$F_2 = M_c(F_1) \otimes F_1 \tag{9}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{10}$$

Here, M_c and M_s represent the channel attention map and the spatial attention map, respectively, while $\otimes$ denotes element-wise multiplication. The channel attention map M_c is generated by the Channel Attention submodule, whose structure is depicted in Fig. 7.

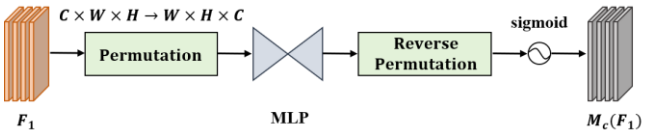


Fig. 7. Channel Attention Submodule.

The Channel Attention submodule consists of four main components: permutation, MLP (Multi-Layer Perceptron), reverse permutation, and sigmoid activation (Fig. 7). First, a 3D permutation is applied to maintain the integrity of information across the three dimensions of the feature map. Next, an MLP is employed to enhance the cross-dimensional channel-space dependency. To optimize performance, the reduction ratio r is used in the MLP. Once the MLP processing is complete, the channel attention map M_c is generated through reverse permutation and a sigmoid activation function.

This design helps preserve information integrity while strengthening the network's ability to handle cross-dimensional features. This enhancement improves the network's efficiency in complex scenarios, allowing it to better capture critical feature dependencies.

The Spatial Attention submodule is designed to improve the network's focus on spatial information and is illustrated in Fig. 8. It consists of two convolutional layers for spatial information fusion, along with a sigmoid activation function. The reduction ratio r, consistent with that in the Channel Attention submodule, is also applied here. Unlike traditional approaches, this submodule removes pooling steps to retain more detailed feature information.

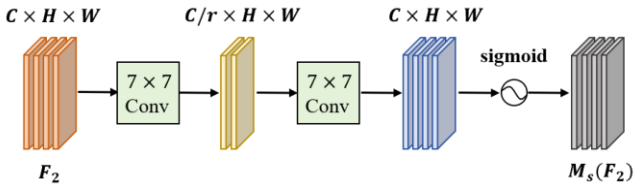


Fig. 8. Spatial Attention Submodule.

While removing pooling may slightly increase the number of parameters, it effectively preserves the completeness of spatial information and reduces the risk of losing key feature mappings. This makes the spatial attention submodule particularly effective in scenarios requiring detailed spatial feature extraction.

By integrating GAM into GMP-YOLO, the loss of critical information caused by increasing network depth is effectively mitigated. GAM enhances the model's ability to process complex information by promoting better global feature interaction and preserving critical details in the feature maps. As a result, the detection performance of GMP-YOLO for remote sensing images is significantly improved.

IV. Analysis of Experiments and Results

This study employs YOLOv8.1.7, developed by

Ultralytics, as the benchmark model. All experimental procedures and model training were conducted within a PyTorch framework (Python 3.9.18 and PyTorch 1.10.0+cu113). The experimental hardware configuration consisted of an Intel(R) Core(TM) i5-12490F CPU and an NVIDIA GeForce RTX 4060 Ti GPU.

Considering the practical deployment constraints of remote sensing applications on edge computing devices, which impose stringent limitations on model parameters, memory utilization, and inference latency, we selected the lightweight YOLOv8-s architecture as our baseline for enhancement. YOLOv8-s maintains the fundamental design principles of the v8 series while being specifically optimized for resource-constrained environments through strategic adjustments to network width and depth.

The critical hyperparameter configurations implemented during model training are detailed in Table I.

TABLE I
TABLE OF TRAINING PARAMETER SETTINGS

Parameters	Setup
Epochs	300
Batch Size	8
Optimizer	SGD
Momentum	0.9
NMS IoU	0.7
Mosaic	1.0

A. Dataset Selection

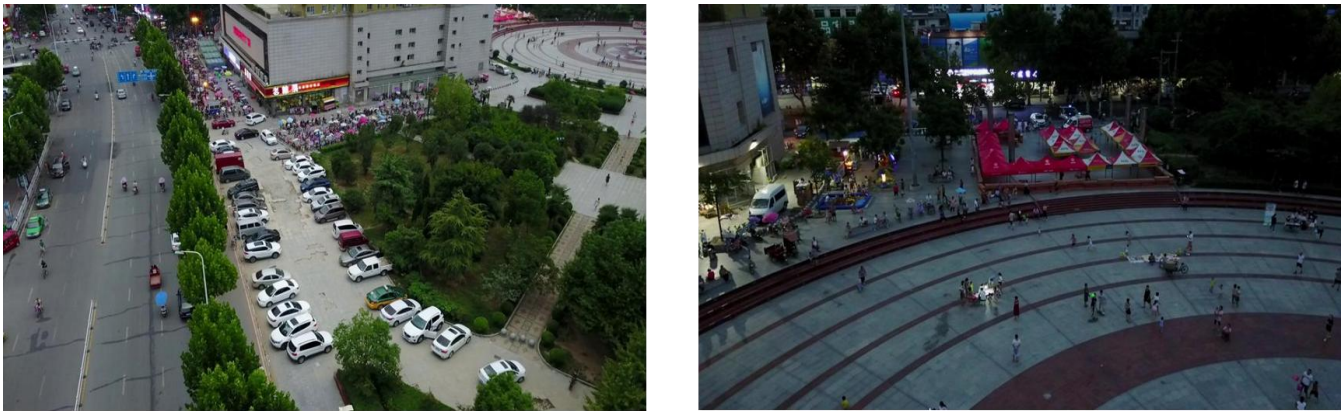
To rigorously evaluate the performance of the proposed GMP-YOLO model, two representative remote sensing image datasets, DOTA and VisDrone [22], were selected. The DOTA dataset, a benchmark in remote sensing image object detection, served as the primary validation platform for ablation experiments due to its scene diversity and comprehensive annotations.

The DOTA (Dataset for Object Detection in Aerial Images) dataset represents one of the most influential resources in remote sensing object detection research. It comprises 2,806 aerial images annotated with 188,282 instances across 15 common categories, as illustrated in Fig. 9. The dataset's complex environmental settings and heterogeneous target distributions create a challenging evaluation scenario, ideal for assessing detection performance. To optimize the dataset for experimental purposes, the original high-resolution images (4000×4000 pixels) were segmented into smaller 1024×1024 pixel subimages. This preprocessing step enhances training efficiency and reduces computational resource requirements. The dataset was systematically partitioned into 15,749 training images, 5,297 validation images, and 12,779 test images to ensure robust performance evaluation. Additionally, the annotation format was converted from the original DOTA specification to the YOLO format to standardize coordinate representations and ensure compatibility with the GMP-YOLO architecture.

The VisDrone2019 dataset constitutes a comprehensive collection of 288 video sequences, encompassing 261,908 video frames and 10,209 static images. Acquired using various drone platforms across diverse geographic locations and environmental conditions, the dataset incorporates varying illumination scenarios and complex scene compositions. Unlike conventional object detection datasets, individual images in VisDrone frequently contain hundreds of targets, with the complete dataset featuring approximately 2.6 million manually annotated bounding boxes. Beyond basic annotations, the dataset provides supplementary metadata regarding scene visibility conditions, target categorization, and occlusion states, rendering it particularly valuable for multi-task applications. Representative examples from this dataset are presented in Fig. 10.

(a) plane **(b) harbor** **(c) ship** **(d) swimming pool**

Fig. 9. Example of a partial image of the DOTA dataset.



(a) Multi-scale objectives (daytime) **(b) Multi-scale objectives (night)**

Fig. 10. Example of a partial image of the VisDrone dataset.

B. Experimental Evaluation Criteria

In this study, we employ several complementary evaluation metrics including precision, recall, mean Average Precision (mAP), and frames per second (FPS) to comprehensively assess the performance of the proposed methodology.

Precision quantifies the model's prediction reliability, representing the proportion of correctly identified targets among all detections. It is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP):

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{11}$$

where TP represents correctly detected targets and FP denotes incorrectly detected targets.

Recall measures the model's detection coverage, indicating the proportion of ground truth targets successfully identified. It is defined as the ratio of true positives (TP) to the sum of true positives and false negatives (FN):

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{12}$$

where FN represents labeled targets that the model failed to detect. These metrics reveal an inherent inverse relationship—improving precision typically results in decreased recall and vice versa. To address this fundamental trade-off and provide a more comprehensive performance assessment, we utilize Average Precision (AP), which integrates precision and recall across varying confidence thresholds. AP is computed by plotting the precision-recall curve and calculating the area under this curve.

For multi-category detection tasks, AP values are calculated independently for each category and subsequently averaged to derive the mean Average Precision (mAP), which serves as a critical metric for holistic model evaluation. The formulations for AP and mAP are expressed as:

$$\text{AP}=\int_0^1 \text{P}(\text{R})\,\text{dr}, \text{mAP}=\frac{1}{\text{N}}\sum_{(\text{i}=1)}^{\text{N}}\text{AP}_\text{i} \tag{13}$$

where N represents the total number of object categories in the detection task, and AP is derived from the precision-recall relationship P(R) across different confidence thresholds.

Additionally, considering the real-time processing requirements inherent in remote sensing applications, we incorporate frames per second (FPS) as a performance indicator to evaluate computational efficiency. The FPS metric quantifies the number of images the model can process per second, providing a crucial benchmark for operational deployability in time-sensitive applications.

C. Ablation Experiment

To systematically evaluate the effectiveness of each proposed component, we conducted comprehensive ablation experiments on the DOTA dataset. Initially, we assessed the impact of the PIoUv2 loss function on model performance. As presented in Table II, the incorporation of PIoUv2 elevated the overall mAP from 66.2% to 68.5%, while simultaneously enhancing computational efficiency, with FPS increasing from 357.1 to 416.7. Although a marginal decrease in small target detection performance was observed, the substantial improvement in overall detection metrics validates the effectiveness of the proposed loss function.

The integration of the Multi-Scale High-Resolution Feature Fusion Network (MSHR-Net) significantly enhanced the model's capacity to detect objects across diverse scale ranges. As detailed in Table III, the AP for small vehicle detection increased from 68.2% to 68.3%, while medium and large targets, exemplified by planes and tennis courts, achieved AP values of 91.7% and 93.8%, respectively. The comprehensive mAP improved from 66.2% to 68.9%, confirming MSHR-Net's efficacy in addressing multi-scale detection challenges. Fig. 11 illustrates the qualitative improvement in detection performance after incorporating MSHR-Net, highlighting enhanced confidence and precision, particularly in complex environmental contexts and multi-target scenarios.

The introduction of the Global Attention Mechanism (GAM) further enhanced detection performance across multiple target categories. As evidenced in Table IV, GAM improved the AP for small vehicle detection to 70.8%, while plane detection AP reached 92.7%. The aggregate mAP across all categories increased substantially from 66.2% to 72.0%, demonstrating GAM's capacity to significantly improve the detection of small and complex targets. Fig. 12 presents a qualitative comparison of detection results, illustrating how the GAM-enhanced model effectively extracts salient features in complex scenes, thereby reducing missed detections, particularly for closely positioned or partially overlapping targets.

TABLE II
ABLATION EXPERIMENTS WITH PIOU V2 (BEST PERFORMANCE HIGHLIGHTED IN BOLD)

Model	Structure	Small-vehicle	Plane	Tennis-court	All Classes			
	PIoU v2	AP(%)			Precision (%)	Recall (%)	mAP (%)	FPS
YOLOv8	-	**68.2**	90.7	93.6	96.7	81.0	66.2	357.1
Ours	✓	66.4	**91.4**	**93.7**	**96.9**	**82.0**	**68.5**	**416.7**

TABLE III
ABLATION EXPERIMENTS WITH MSHR-NET (BEST PERFORMANCE HIGHLIGHTED IN BOLD)

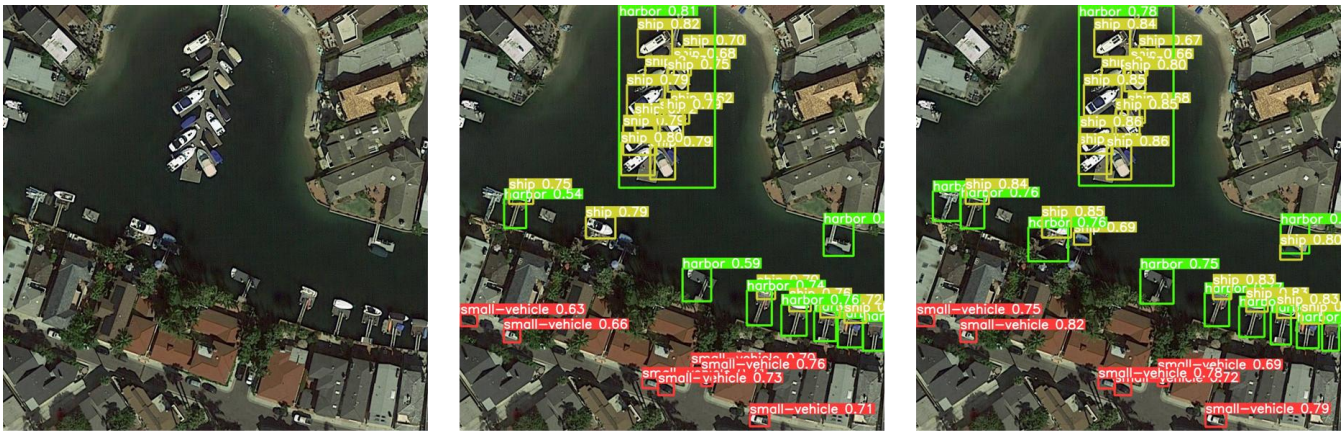
Model	Structure	Small-vehicle	Plane	Tennis-court	All Classes		
	MSHR-Net	AP(%)			Precision (%)	Recall (%)	mAP (%)
YOLOv8	-	68.2	90.7	93.6	**96.7**	81.0	66.2
Ours	✓	**68.3**	**91.7**	**93.8**	94.8	**81.0**	**68.9**

(a) (b) (c)

Fig. 11. Comparison of detection results. (a) Original image. (b) YOLOv8. (c) Model after MSHR-Net addition.

TABLE IV
ABLATION EXPERIMENTS WITH GAM (BEST PERFORMANCE HIGHLIGHTED IN BOLD)

Model	Structure	Small-vehicle	Plane	Tennis-court	All Classes		
	GAM	AP(%)			Precision (%)	Recall (%)	mAP (%)
YOLOv8	-	68.2	90.7	93.6	**96.7**	81.0	66.2
Ours	✓	**70.8**	**92.7**	**95.0**	95.1	**82.0**	**72.0**



(a) (b) (c)

Fig. 12. Comparison of detection results. (a) Original image. (b) YOLOv8. (c) GAM-added model.

TABLE V
ABLATION EXPERIMENTS OF THE PROPOSED MODULES ON THE DOTA DATASET (BEST PERFORMANCE HIGHLIGHTED IN BOLD)

Model	Structure			Small-vehicle	Plane	Tennis-court	All Classes	
	PIoU v2	MSHR	GAM	AP(%)			mAP(%)	FPS
YOLOv8	-	-	-	68.2	90.7	93.6	66.2	357.1
	✓	-	-	66.4	91.4	93.7	68.5	**416.7**
	-	✓	-	68.3	91.7	93.8	68.9	285.7
Ours	-	✓	✓	**70.9**	92.4	95.1	72.1	277.8
	✓	✓	-	68.5	91.8	94.6	69.1	333.4
	✓	✓	✓	70.5	**92.7**	**95.2**	**73.2**	303.0

Finally, we conducted integrated ablation experiments to evaluate the combined effect of all proposed modules in the GMP-YOLO architecture. As detailed in Table V, optimal performance was achieved when PIoUv2, MSHR-Net, and GAM were simultaneously implemented. The aggregate mAP improved to 73.2%, representing a substantial 7.0 percentage point increase over the baseline, while maintaining a competitive FPS of 303.0. These results demonstrate the complementary nature of the proposed modules, which collectively enhance the model's overall detection capabilities.

Through these systematic ablation studies, we have validated the contribution of each architectural enhancement to the model's overall performance. The PIoUv2 loss function provides more precise bounding box regression, MSHR-Net enhances multi-scale target detection capabilities, while GAM improves feature extraction through its attention mechanism. The synergistic integration of these three components enables GMP-YOLO to achieve significant improvements in detection accuracy while maintaining efficient computational performance.

D. Generalized Performance Verification

To rigorously assess the generalization capabilities of the proposed GMP-YOLO architecture, we conducted extensive experiments on the VisDrone dataset, which presents exceptional diversity and complexity. This dataset encompasses multiple object categories — including pedestrians and various vehicle types—providing an authentic representation of practical detection scenarios in urban, low-altitude environments. The comparative experimental results are summarized in Table VI, with optimal performance metrics highlighted in bold.

For pedestrian detection, GMP-YOLO achieved an AP of 42.0%, representing a substantial improvement over the baseline YOLOv8 model's 35.8%. Across vehicle categories (cars, vans, and buses), GMP-YOLO consistently outperformed the baseline, attaining AP values of 79.0%, 44.2%, and 60.1%, respectively. Particularly noteworthy is the 15 percentage point improvement in bus detection, demonstrating significantly enhanced performance in complex traffic environments. Additionally, GMP-YOLO exhibited superior detection capabilities for motorized vehicles with an AP of 44.8%, compared to YOLOv8's 37.0%. The aggregate mAP across all categories reached 39.7% for GMP-YOLO—a substantial 6.5 percentage point improvement over YOLOv8's 33.2%—comprehensively validating its enhanced performance in multi-category object detection tasks.

Fig. 13 presents representative detection examples from the VisDrone dataset under diverse conditions. In both daylight (Fig. 13(a)) and nocturnal (Fig. 13(b)) environments, GMP-YOLO demonstrated exceptional performance. In daylight scenarios, the model successfully identified and precisely localized various objects, maintaining consistent performance despite complex background elements. In nocturnal scenarios, despite the inherent challenges of limited illumination and increased background noise, GMP-YOLO effectively completed detection tasks with minimal false negatives or false positives. These examples provide compelling evidence of GMP-YOLO's robustness across varying illumination conditions and environmental contexts, highlighting its broad applicability in complex urban scenarios.

Based on its superior performance on the VisDrone dataset, we selected this domain-adapted model as the foundation for subsequent intelligent agent development. The VisDrone dataset not only encompasses diverse urban targets—including pedestrians, various vehicle types, and infrastructure elements—but also presents significant challenges under varying illumination conditions and complex environmental contexts. This comprehensive validation underscores GMP-YOLO's reliability and generalization capacity in dynamic and intricate environments. Leveraging this optimized model, we developed an urban planning analysis intelligent agent capable of real-time object identification and spatial analysis in low-altitude urban environments. Through precise localization and classification, the agent provides critical data support for urban planning and management applications, thereby facilitating dynamic, evidence-based decision-making processes.

TABLE VI
EXPERIMENTAL RESULTS ON VISDRONE DATASET (BEST PERFORMANCE HIGHLIGHTED IN BOLD)

Model	Pedestrian	Car	Van	Bus	Motor	All Classes
	AP(%)					mAP(%)
YOLOv8	35.8	76.0	39.2	45.1	37.0	33.2
Ours	**42.0**	**79.0**	**44.2**	**60.1**	**44.8**	**39.7**

(a) (b)

Fig. 13. Example of detection results in VisDrone dataset. (a) Daytime. (b) Night.

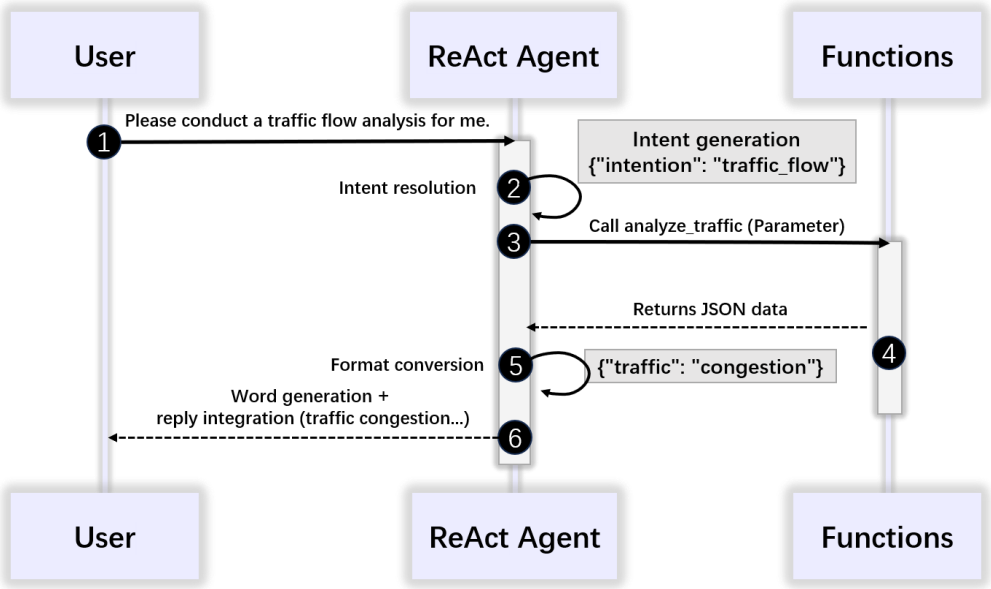


Fig. 14. ReAct Agent Decision Process.

V. DESIGN AND IMPLEMENTATION OF THE INTELLIGENT AGENT

In response to the rapid expansion of the low-altitude economy and the increasing demand for intelligent urban management, we have developed a comprehensive smart city remote sensing image analysis system based on our GMP-YOLO architecture. This system specifically addresses the analytical challenges presented by urban low-altitude environments, where detailed object detection and contextual interpretation are critical for effective decision-making.

A. System Architecture and Analytical Capabilities

The system architecture utilizes the GMP-YOLO model as its core analytical engine, enhancing detection capabilities for low-altitude remote sensing imagery. Based on this foundation, we have developed four specialized analytical tool modules for agent utilization: Traffic Flow Analysis Tool processes vehicle and pedestrian data to calculate real-time traffic conditions; Natural Environment Assessment Tool evaluates urban environmental quality; Infrastructure Evaluation Tool identifies and assesses urban infrastructure elements; and Emergency Event Detection Tool identifies abnormal patterns for rapid response.

These specialized analytical tools are coordinated by a ReAct (Reasoning and Acting) agent built on the LangChain and LangGraph frameworks. This agent architecture, illustrated in Fig. 14, enables dynamic tool selection and sequential task execution based on specific analytical requirements and contextual understanding.

The ReAct agent implementation provides several key advantages for urban image analysis:

1) Context-Aware Processing: The agent maintains session context through unique conversation identifiers, enabling coherent multi-round interactions and progressive analytical refinement.
2) Dynamic Tool Selection: Based on user queries and image content, the agent intelligently determines which analytical tools to deploy and in what sequence, optimizing computational efficiency.
3) Integrated Reasoning: The agent combines detection results with contextual knowledge to generate comprehensive analytical reports that extend beyond simple object detection to meaningful urban insights.

B. System Workflow and User Interaction

The operational workflow begins with user authentication through the login interface shown in Fig. 15. Users can either upload custom imagery or select from pre-established datasets for analysis.

Upon image submission (Fig. 16), the front-end transmits the data to the back-end processing pipeline, where the ReAct agent parses the analytical requirements and activates the GMP-YOLO model for initial object detection.

The detection results, including object categories, spatial coordinates, and confidence metrics, are then processed by the specialized analytical tools selected by the agent. These tools generate detailed analytical outputs, which are compiled into a comprehensive report featuring annotated imagery, quantitative metrics, and actionable insights generated through large language model interpretation.

The final results are presented through an interactive web interface (Fig. 17), allowing users to explore different aspects of the analysis and download complete reports. Additionally, users can engage in follow-up inquiries to obtain more specific information or predictive insights based on the detected patterns.

This intelligent agent system demonstrates how advanced object detection models like GMP-YOLO can be effectively integrated into comprehensive urban management solutions. By combining precise detection capabilities with specialized analytical tools and an intelligent coordination mechanism, the system provides valuable decision support for urban planning, traffic management, environmental monitoring, and emergency response applications in the rapidly evolving low-altitude urban environment.

Fig. 15. Login Interface.

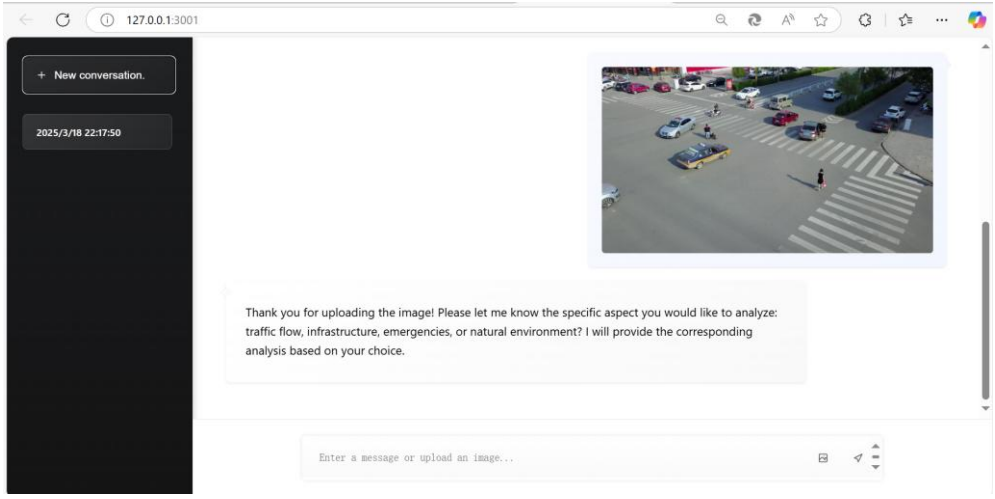


Fig. 16. Uploaded Image.

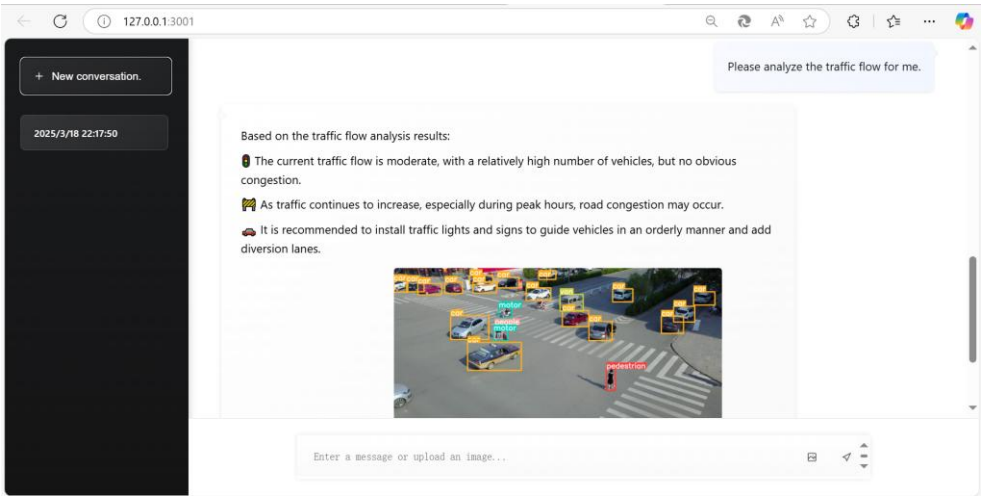


Fig. 17. Results Return Interface.

VI. Conclusion

This study presents GMP-YOLO, a novel approach to multi-scale object detection in remote sensing imagery that addresses significant challenges in this domain. Conventional object detection methodologies continue to exhibit limitations in accurately detecting targets across diverse scales and adapting to complex environmental contexts. To overcome these constraints, we have developed GMP-YOLO, a comprehensive detection framework that demonstrates exceptional performance in remote sensing imagery through advanced feature extraction and fusion strategies.

From an architectural perspective, GMP-YOLO integrates a Multi-Scale High-Resolution Feature Fusion Network (MSHR-Net) with a specialized Multi-scale Detection Block (MDB). The model incorporates dedicated detection scales for tiny targets, substantially improving detection accuracy across objects of varying dimensions. For bounding box localization, we replaced the conventional CIoU loss function with an enhanced PIoUv2 loss function, which significantly improves anchor box regression precision while accelerating model convergence. Additionally, we implemented a Global Attention Mechanism (GAM) to enhance the extraction and integration of salient features.

Comprehensive evaluations on the DOTA and VisDrone datasets demonstrate that GMP-YOLO consistently outperforms established baseline models in both detection accuracy and generalization capabilities. The model exhibits particularly strong performance in challenging scenarios involving complex backgrounds and small targets. Furthermore, GMP-YOLO maintains robust stability under adverse conditions, including intricate traffic environments and variable illumination, substantially reducing both false positive and false negative detections.

Building upon these algorithmic advancements, we designed and implemented a complete remote sensing image analysis system with GMP-YOLO as its foundational detection engine. This system represents a significant engineering contribution, featuring an intelligent agent architecture capable of real-time object identification and contextual analysis in urban environments. We developed this system by integrating the LangChain framework and implementing the ReAct Agent mechanism, enabling dynamic coordination of multiple specialized analytical tools based on specific task requirements. Our comprehensive development effort included creating four specialized analytical modules: object detection, traffic flow analysis, environmental quality assessment, and infrastructure evaluation. These components work in concert to deliver unprecedented accuracy and efficiency in urban remote sensing analysis applications.

This research not only validates GMP-YOLO's effectiveness for remote sensing object detection but also demonstrates how advanced detection models can be transformed into functional intelligent systems through careful system design and integration. Future research directions include optimizing computational efficiency and inference speed, particularly for deployment on resource-constrained edge computing devices. These optimizations will further enhance support for real-time remote sensing applications and extend the practical utility of our model and system across intelligent urban management and remote sensing domains.

References

[1] T. Wellmann et al., “Remote Sensing in urban planning: Contributions towards ecologically sound policies?,” Landscape and Urban Planning, vol. 204, p. 103921, Dec. 2020.

[2] Z. Zou and Z. Shi, “Random Access Memories: A new paradigm for target detection in high resolution aerial remote sensing images,” IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1100–1111, Mar. 2018.

[3] J. Zhan, Y. Hu, W. Cai, G. Zhou, and L. Li, “PDAM–stpnnet: A small target detection approach for wildland fire smoke through remote sensing images,” Symmetry, vol. 13, no. 12, p. 2260, Nov. 2021.

[4] R. Girshick, “Fast R-CNN,” 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440–1448, Dec. 2015.

[5] W. Liu et al., “SSD: Single shot multibox detector,” Lecture Notes in Computer Science, pp. 21–37, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788, Jun. 2016.

[7] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020.

[8] Y. Li, Q. Fan, H. Huang, Z. Han, and Q. Gu, “A modified Yolov8 Detection Network for UAV aerial image recognition,” Drones, vol. 7, no. 5, p. 304, May 2023.

[9] A. Van Etten, “You only look twice: Rapid multi-scale object detection in satellite imagery,” arXiv preprint arXiv:1805.09512, 2018.

[10] Z. Huang et al., “An improved method for ship target detection based on yolov4,” Applied Sciences, vol. 13, no. 3, p. 1302, Jan. 2023.

[11] Z. Du and Y. Liang, “Object Detection of Remote Sensing Image Based on Multi-Scale Feature Fusion and Attention Mechanism,” IEEE Access, 2024.

[12] P. Zhang et al., "Research on improved lightweight Yolov5s for multi-scale ship target detection," Applied Sciences, vol. 14, no. 14, p. 6075, Jul. 2024.
[13] L.-S. Wei, S.-H. H, and L.-Y. M, "MTD-YOLOv5: Enhancing marine target detection with multi-scale feature fusion in YOLOv5 model," Heliyon, vol. 10, no. 4, 2024.
[14] C. Liu, K. Wang, Q. Li, et al., "Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism," Neural Networks, vol. 170, pp. 276–284, 2024.
[15] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," Lecture Notes in Computer Science, pp. 404–419, 2018.
[16] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," arXiv preprint arXiv:2112.05561, 2021.
[17] C. Y. Wang, H. Y. M. Liao, and I. H. Yeh, "Designing network design strategies through gradient path analysis," arXiv preprint arXiv:2211.04800, 2022.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[19] Z. Ge, S. Liu, F. Wang, et al., "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.
[20] Z. Zheng et al., "Distance-IOU loss: Faster and better learning for bounding box regression," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12993–13000, Apr. 2020.
[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018.
[22] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018.