Object Detection Model for Remote Sensing Images Based on YOLOv9

Donghao Hou, Yujun Zhang*

***Abstract*—In the field of object detection for remote sensing images, especially in applications such as environmental monitoring and urban planning, significant progress has been made. This paper addresses the common challenges faced by traditional object detection methods in remote sensing images, such as the large number of targets and complex backgrounds, by proposing a novel network based on YOLOv9. The network innovatively introduces the C3_CD_CGA module, an enhanced module based on Cascaded Group Attention, designed to reduce computational redundancy and increase attention diversity, and enhances the processing capability of multi-scale information through the CD module. The C3 module employs deep asymmetric convolution to mitigate information loss and increase the receptive field. Additionally, the network integrates DSConv with the RepNCSPELAN4 module to adaptively focus on and precisely capture the features of elongated and curved local structures, such as vehicles. The introduction of the CARAFE module further improves the spatial resolution of the feature maps, significantly enhancing performance across various visual tasks. Experimental results show that the improved YOLOv9 achieves a mean average precision (mAP) of 88% on the SIMD dataset, which is an improvement of 1.6% compared to the baseline YOLOv9 model and 1.5% higher than the state-of-the-art YOLO-SE model. This model not only achieves more effective multi-target recognition in complex backgrounds but also strikes a good balance between accuracy and efficiency.**

***Index Terms*—Attention mechanism, Object detection, Remote sensing images, YOLOv9.**

I. Introduction

In recent years, with the rapid development of remote sensing technology and the large-scale acquisition of satellite image data, the application areas of remote sensing images have continuously expanded [1–4], covering important fields such as urban planning, environmental monitoring, disaster management, and agricultural monitoring. Object detection technology not only needs to accurately identify and locate various objects in images but also must address challenges related to different scales, complex environmental conditions, and diverse target shapes [5–7]. Traditional remote sensing object detection methods often rely on manually designed feature extractors and classifiers, which face limitations in accuracy and generalization when dealing with large-scale, diverse datasets. With the rise of deep learning, particularly Convolutional Neural Networks (CNNs), data-driven end-to-end object detection methods have significantly improved detection accuracy and efficiency. Among these, the You Only Look Once (YOLO) series of models has gained widespread attention for its real-time processing capabilities and high accuracy. As an evolved version of the YOLO series, YOLOv9 further enhances object detection capabilities in complex scenarios by introducing more sophisticated network architectures and advanced training strategies [8]. However, despite its excellent performance in general visual scenes, YOLOv9 still faces challenges when applied to specific remote sensing datasets, such as the Satellite Image Multi-Scale Detection (SIMD) dataset.

Wang et al. introduced a novel method based on Graph Neural Networks (GNNs) to address the challenges of complex target association and diverse target categories. GNNs can effectively model the relationships between targets, thereby improving the accuracy and efficiency of object detection. By constructing a relational graph between targets, the model can better understand and utilize the association information among targets, enhancing its capability to detect multiple target categories [9].Shen et al. proposed a new optimization method focused on handling the issue of multiple target categories. By introducing a dynamic weight allocation mechanism, this method automatically adjusts the weights of each category during training based on the number of samples and the difficulty of detection for each category. This balances the training process across categories, thereby improving the overall performance of the detection model [10].Xu et al. presented a data augmentation technique for large-scale category detection that enhances the generalization capability and detection accuracy of detectors by generating diverse target samples using Generative Adversarial Networks (GANs). This approach is particularly effective for long-tail distribution detection tasks, as it generates more samples for minority classes, balancing the data distribution and significantly improving the model's ability to detect minority targets [11].Zhang et al. proposed a method that combines Generative Adversarial Networks (GANs) with reinforcement learning to improve object detection. They utilized GAN generators to create more realistic small target samples, aiding the detector in better recognizing and locating small targets. Reinforcement learning was further employed to optimize the detector, expanding the model's receptive field and adapting it to different target categories, thereby enhancing overall detection performance [12].Liu et al. explored an object detection method based on multi-scale feature fusion and dilated convolutional networks in their research. Dilated convolutional networks expand the receptive field, enabling the model to capture more contextual information, thus improving the detection accuracy of small targets [13]. Additionally, the multi-scale feature fusion technique allows the model to handle targets of different scales simultaneously, enhancing its ability to detect diverse targets. This method performs excellently in various

Manuscript received September 4, 2024; revised January 24, 2025. This work was supported by the Key Laboratory of Internet of Things Application Technology on Intelligent Construction, Liaoning Province (2021JH13/10200051)

Donghao Hou is a graduate student of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 3070542820@qq.com).

Yujun Zhang is a Professor of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding Author, e-mail: 1997zyj@163.com).

complex scenarios, particularly in detecting small targets and handling complex backgrounds, where detection results are significantly improved.

The improved YOLOv9 model performs better in handling multi-scale, multi-category, and complex environment object detection tasks, while significantly enhancing overall performance.

The improved YOLOv9 model proposed in this paper incorporates the C3_CD_CGA module into its neck network. The C3 module utilizes multi-scale feature fusion techniques to enhance the model's feature extraction capabilities across different scales, and employs dilated convolutions to expand the receptive field, thereby capturing more contextual information. The CGA module optimizes the representation of multi-scale features through intra-group and inter-group attention mechanisms, providing feature subsets to different attention heads, reducing computational redundancy, and increasing attention diversity. By combining the C3, CD, and CGA modules [14, 15], the model achieves multi-scale feature fusion and addresses the challenge of target diversity.

In the backbone network, the DSConv module is combined with the RepNCSPELAN4 module [16]. The DSConv module adopts a multi-view feature fusion strategy to enhance the focus on features from multiple perspectives, ensuring that key information from different global shapes is preserved during the feature fusion process. In the neck network, the upsampling module is replaced with the CARAFE module [17], which significantly improves the spatial resolution of feature maps through adaptive interpolation and reassembly techniques. The CARAFE module, by leveraging an automatically learned reassembly process, more effectively retains and reconstructs information within feature maps, capturing contextual information from different locations. This not only enhances the model's performance but also maintains computational efficiency, effectively addressing the challenges posed by complex environmental conditions. Experimental results on the SIMD dataset validate the effectiveness of this approach.The specific contributions are as follows:

The SIMD dataset is characterized by data imbalance and diverse target categories. An attention module was designed to address the issue of target diversity, significantly enhancing overall performance while reducing computational complexity.

Due to the varying image scales and resolutions, as well as complex environmental conditions in the SIMD dataset, a multi-view feature fusion strategy was employed. This approach significantly improves the spatial resolution of feature maps and adapts to complex environmental conditions.

II. Related work

In recent years, YOLOv9, as the latest model in the YOLO series, has demonstrated powerful performance in handling aerial imagery through its improved network architecture and training strategies. YOLOv9 has further enhanced the model's detection accuracy, speed, and ability to handle multi-scale targets by comprehensively optimizing the backbone, neck, and head networks. Compared to previous YOLO models, YOLOv9 exhibits a strong competitive advantage.

Compared to YOLOv3 and YOLOv4, YOLOv9 backbone network incorporates a more advanced deep network structure, combining Cross-Stage Partial Network (CSPNet) technology with improved residual connections [18]. These enhancements significantly boost the model's feature extraction capabilities, enabling it to more effectively capture diverse image features in complex aerial imagery, particularly excelling in scenarios involving small targets and detail-rich scenes. While YOLOv5 also adopts the CSP architecture, YOLOv9 backbone network further advances the richness of feature representation and processing capabilities, offering higher detection accuracy.

The neck of YOLOv9 incorporates an improved Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) [19, 20], significantly enhancing the fusion and transmission of multi-scale features. This design enables YOLOv9 to achieve greater precision when handling multi-scale targets, particularly in aerial imagery, where the size and shape of objects often vary greatly. The optimization of the neck network allows YOLOv9 to maintain excellent detection performance across different scales. Compared to YOLOv4, YOLOv9 offers more refined multi-scale fusion, providing better robustness and detection performance in the presence of complex backgrounds and densely packed targets.

By introducing a self-attention mechanism and optimized activation functions, YOLOv9 achieves higher levels of accuracy and speed in object detection. The application of the self-attention mechanism allows YOLOv9 to more precisely identify targets in complex scenes, reducing false positives and missed detections. Additionally, the new loss function design further optimizes the regression accuracy of bounding boxes, resulting in more accurate localization in high-resolution aerial imagery. Compared to the head network design of YOLOv5, YOLOv9 demonstrates superior performance in handling complex backgrounds and small targets, making it more advantageous in practical applications.

III. Method Introduction

A. Modules of the Improved YOLOv9 Algorithm

While YOLOv9 can improve classification accuracy by optimizing the classification loss function when handling multi-category targets, its classification performance may decline when the number of categories is large. This is because the feature differences between categories are relatively small, making it challenging for the model to distinguish between similar categories. Additionally, the YOLOv9 model has limited ability to detect objects in complex backgrounds, particularly in the SIMD dataset, where the complex backgrounds in drone-captured images can easily interfere with the model's detection results. Therefore, although YOLOv9 performs excellently in some scenarios, further improvements and optimizations are still needed to effectively handle complex and variable drone datasets.

To address these challenges, this paper proposes an enhanced structure for YOLOv9. First, the Upsample module was replaced with the CARAFE module, which introduces a novel information reorganization mechanism, significantly improving the spatial resolution of feature maps. Second, the backbone network integrates the RepNCSPELAN4 and DSConv modules to form the DSConvRepNCSPELAN4 module, greatly enhancing YOLOv9 capability in multi-scale object detection. This enhancement is particularly effective

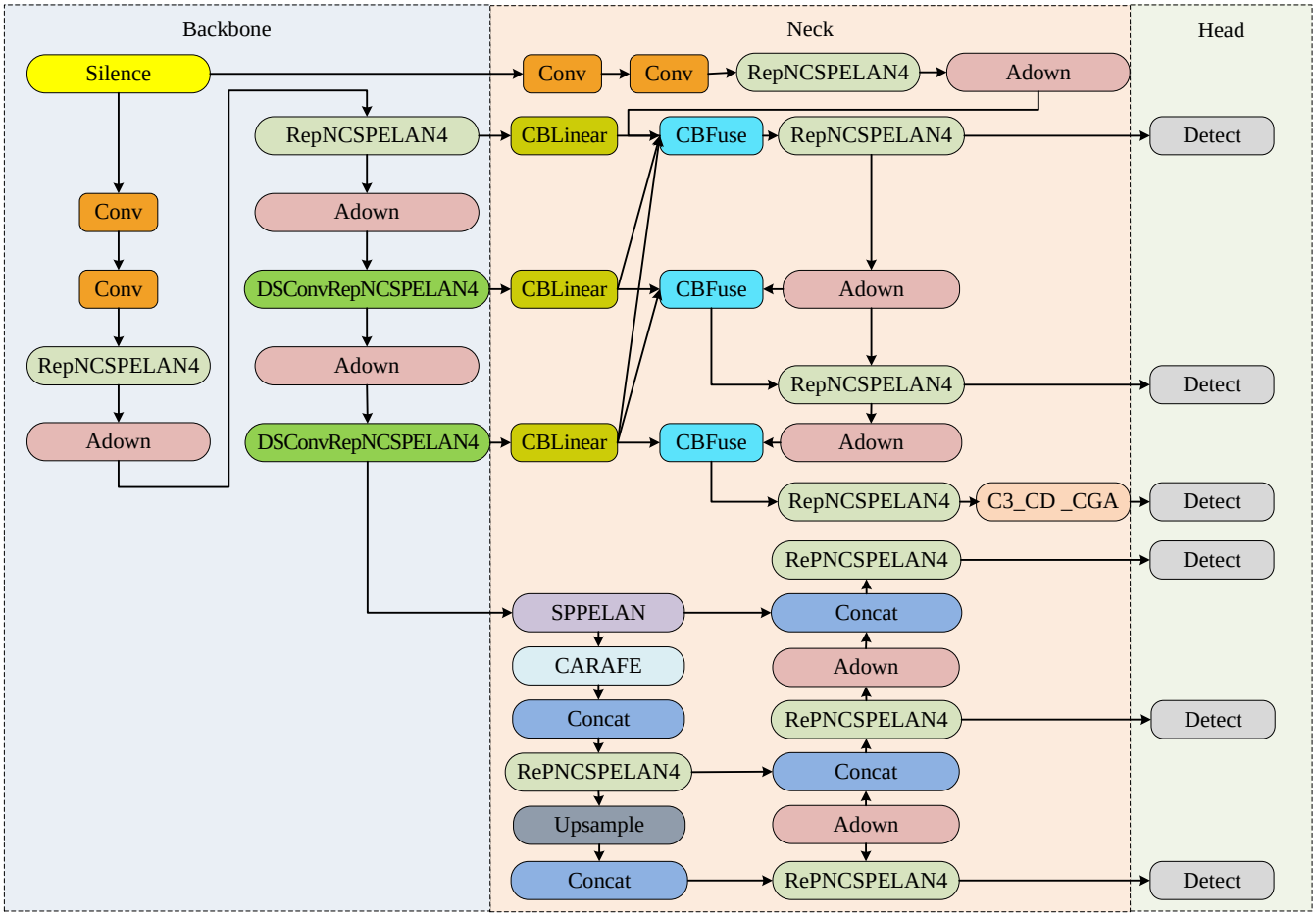

Fig. 1: Overall improved architecture diagram

in accurately identifying and locating targets of different scales in drone images from the SIMD dataset. Finally, to improve the accuracy of multi-category object detection, the integration of the C3_CD_CGA module not only increases YOLOv9 accuracy in multi-category, multi-scale object detection but also demonstrates stronger robustness in complex scenarios, thereby comprehensively improving YOLOv9 detection performance.

B. C3_CD_CGA

Integrating the C3_CD_CGA module into the neck of YOLOv9 enhances the capability to capture diverse classes and multi-scale image features. To ensure the completeness of information during image processing, the C3_CD_CGA module incorporates the C3 (Concentrated-Comprehensive Convolution) module. The C3 module applies dilated convolutions to depthwise separable convolutions, as described in Equation 8, where K^d denotes the depth-wise convolution kernel, d represents the dilation rate, and K^p stands for the 1×1 point-wise convolution kernel. The C3 module processes features through two stages: the first stage is the concentration stage, which utilizes depth-wise separable asymmetric convolutions to capture information from neighboring pixels, thereby alleviating local information loss caused by dilated convolutions. The second stage is the comprehensive convolution stage, which employs depth-wise separable dilated convolutions to expand the receptive field, while simultaneously using point-wise convolutions to mix channel information. This approach effectively integrates local and global information, reducing parameters and computational complexity while maintaining the model's performance in semantic segmentation.

$$\begin{aligned} F'_{c,h,w} &= \sum_m \sum_n F_{c,h+dm,w+dn} K^d_{c,m,n} \\ O_{c',h,w} &= \sum_c F'_{c,h,w} K^p_{c',c} \end{aligned} \tag{1}$$

The CD convolution in the C3_CD_CGA module uses depth-wise dilated convolution layers with a dilation rate of 2. This dilation technique expands the receptive field of the feature map without increasing the number of parameters and computational complexity, allowing the feature map to cover a larger input area while maintaining its resolution. This enables the network to capture richer contextual information. By combining dilated convolution with the CGA module, the feature extraction capability is further enhanced, strengthening the long-range dependencies between features. This allows the model to achieve a larger receptive field and better global information integration when processing images, thereby improving the model's performance in handling complex scenes.

To enhance the global integration capability of features, the CGA (Cascaded Group Attention) module employs a multi-head self-attention mechanism, allowing for multiple independent feature interactions. The role of the CGA module is to process input features by providing different feature

splits for each attention head, with each head focusing on only a portion of the input. The outputs of all heads are then cascaded together and passed through a linear projection layer to form the final output, unlike traditional self-attention mechanisms that provide the same full feature set to all heads. This approach not only reduces computational redundancy in multi-head attention but also increases the diversity of attention by introducing different features to each attention head.This attention module is shown in Equation 9, where the j attention head performs self-attention on X_{ij}. Here, X_{ij} is the j part of the input feature X_i, X_i is divided into $[X_{i1}, X_{i2}, \ldots, X_{ih}]$, where $1 \leq j \leq h$ is the number of attention heads. The projection layer $W_{ij}^Q, W_{ij}^K, W_{ij}^V$ maps the input features into different subspaces by dividing them into chunks. Subsequently, a linear layer W_i^P re-projects the concatenated output features to match the input dimensions. This design allows the model to capture features at different levels, enhancing the interaction between features through cascading, while improving computational efficiency and the model's ability to capture different subsets of features.

$$\begin{aligned} \tilde{X}_{ij} &= \text{Attn}(X_{ij}W_{ij}^Q, X_{ij}W_{ij}^K, X_{ij}W_{ij}^V) \\ \tilde{X}_{i+1} &= \text{Concat}[\tilde{X}_{ij}]_{j=1:h}W_i^P \end{aligned} \tag{2}$$

In the enhancement of the C3_CD_CGA module within the YOLOv9 model, efforts have been made to maximize the receptive field without significantly increasing the computational burden. Additionally, redundancy issues in multi-head attention have been addressed to ensure that the model accurately identifies and distinguishes targets in complex scenarios. This improvement optimizes the interactions between channels, thereby enhancing the overall capacity and efficiency of the model.

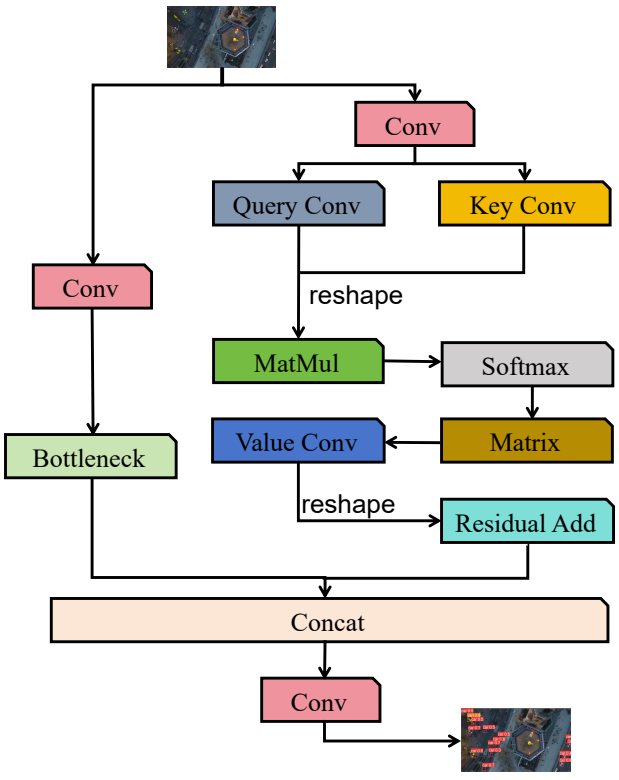

Fig. 2: C3_CD_CGA network architecture

C. DSConvRepNCSPELAN4

The SIMD dataset contains many images with complex environments. In YOLOv9, the RepNCSPELAN4 module enhances the breadth and depth of feature extraction by integrating channel attention and multi-scale feature representation. The DSConv module, based on dynamic snake convolution, can adaptively focus on elongated and winding local features, making it particularly suitable for handling complex tubular structures. The formula for calculating the center position coordinates of the DSConv convolution kernel is given in Equation 5. This module adjusts the shape and receptive field of the convolution kernel to better adapt to the geometric structure of the target, thereby enhancing the model's ability to perceive fine structures.

$$K_i = (x_i, y_i) \tag{3}$$

The formula for unrolling the standard convolution kernel along the x-axis is shown in Equation 6, and the formula for unrolling the standard convolution kernel along the y-axis is shown in Equation 7, where x_i and y_i represent the positions of the convolution kernel in the image coordinates. DySnake convolution adjusts the offset Δ at each position through an iterative strategy, ensuring that the convolution kernel adapts to the target's shape.

$$K_{i\pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = (x_i + c, y_i + \sum_i^{i+c} \Delta y) \\ (x_{i-c}, y_{i-c}) = (x_i - c, y_i + \sum_{i-c}^{i} \Delta y) \end{cases} \tag{4}$$

$$K_{j\pm c} = \begin{cases} (x_{j+c}, y_{j+c}) = (x_j + \sum_j^{j+c} \Delta x, y_j + c) \\ (x_{j-c}, y_{j-c}) = (x_j + \sum_{j-c}^{j} \Delta x, y_j - c) \end{cases} \tag{5}$$

The DSConvRepNCSPELAN4 module combines the precise capture capabilities of DSConv for local complex structures with the multi-scale global feature extraction abilities of RepNCSPELAN4, enabling YOLOv9 to more accurately locate and recognize targets in remote sensing image. This combination significantly enhances the model's detection accuracy across different environments and backgrounds, while maintaining model lightweightness and greatly improving its ability to handle targets of various scales in complex backgrounds, thereby reducing instances of missed and false detections.The DSConvRepNCSPELAN4 module first utilizes DSConv to focus on the key geometric features of slender targets, and then employs the multi-scale feature fusion strategy of RepNCSPELAN4 to further enhance the model's understanding and detection of the overall shape of the targets. Initially, the input feature map undergoes preliminary feature extraction through a convolutional layer, and then it is split into two parts for parallel processing. One part passes through the RepNCSPE module, which enhances feature extraction capabilities, especially for detecting irregular shapes. The other part goes through the DSConv module, which is designed to handle complex shapes like tubular structures by dynamically adjusting convolutional kernels to better capture key features. Additionally, some feature maps directly pass through a convolutional layer, providing additional feature information. Finally, all processed feature maps are concatenated and passed through a final convolutional layer to produce the output.This integration enables the model to exhibit higher precision and robustness when dealing with complex scenes, multiple categories, and multi-scale targets.

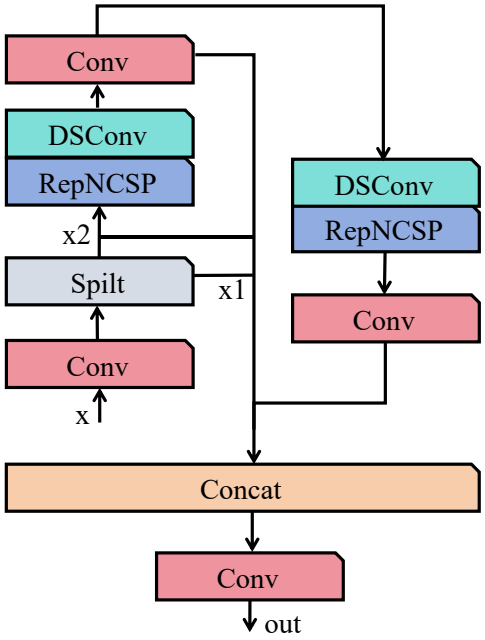

Fig. 3: DSConvRepNCSPELAN4 network architecture

D. CARAFE

Due to the presence of a large number of low-resolution images in the SIMD UAV dataset, we chose to replace the Upsample module with the CARAFE module. The core innovation of CARAFE lies in its ability to dynamically predict the reassembly kernel based on the content of the input features, and to reassemble the features of local regions through weighted combinations. Unlike traditional upsampling methods such as bilinear interpolation or transposed convolution, the reassembly kernel in CARAFE is not fixed, but is dynamically generated at each location by a content encoder. The feature reassembly operation in CARAFE is performed through a weighted summation formula, as shown in Equation 1. In this formula, for the target position l' and the square region centered on l = (i, j) with a size of $N(X_l, k_{\text{up}})$, $r = \left\lfloor \frac{k_{\text{up}}}{2} \right\rfloor$ and $W_{l'}(n, m)$ are reassembly kernels predicted by the content encoder, which determine the contribution of each pixel in the local region $N(X_l, k_{\text{up}})$ to the upsampled pixel l'. This weighted summation method ensures that the reassembled feature map better preserves spatial structure and semantic information.

$$X'_{l'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W_{l'}(n, m) \cdot X(i+n, j+m) \quad (6)$$

CARAFE serves as a reassembly operator with a content-aware kernel. It includes two steps. The first step is to predict a reassembly kernel based on the content of each target location, as shown in Equation 2. For a given input feature map X, CARAFE first predicts the convolution kernel $W_{l'}$ for each target location l', where ψ is the kernel prediction module. The second step is to reassemble the features using the predicted kernel, as shown in Equation 3, where ϕ is the content-aware reassembly module, and the feature $N(X_l, k_{\text{up}})$ is reassembled using the convolution kernel $W_{l'}$.

$$W_r = \psi\left(N\left(X_l, k_{\text{encoder}}\right)\right) \quad (7)$$

$$X'_l = \phi\left(N\left(X_l, k_{\text{up}}\right), W_r\right) \quad (8)$$

Although CARAFE introduces a mechanism for dynamically generating reassembly kernels, it maintains extremely high computational efficiency. By incorporating a channel compressor and a well-designed convolution kernel, CARAFE achieves significant performance improvements with minimal computational overhead, making it easy to integrate into modern neural network architectures. The computational complexity of CARAFE is described by the formula in Equation 4, where C_{in} represents the number of channels in the input feature map, C_m is the number of channels after channel compression, K_{encoder} and K_{up} are the sizes of the convolution kernels for the content encoder and reassembly kernel, respectively, and σ is the upsampling factor.

$$\begin{aligned} FLOPs =& 2(C_{in}+1)C_m + 2(C_m k_{encoder}^2 + 1)\sigma^2 k_{up}^2 \\ &+ 2\sigma^2 k_{up}^2 C_{in} \end{aligned} \quad (9)$$

CARAFE is designed with a large receptive field feature aggregation method that can aggregate contextual information over a wide range. This enables the model to better capture the relationship between local features and global semantics, resulting in excellent performance in tasks such as object detection and semantic segmentation.

IV. Experimental Design and Implementation

A. Dataset Introduction

To validate the effectiveness of the improvements to the YOLOv9 algorithm, this study utilized the Satellite Imagery Multi-Vehicle Dataset (SIMD) and an adapted single-channel deep multi-scale object detection framework, aimed at detecting multi-size/type objects to meet the needs of vehicle ground perspective. The dataset images were obtained from various locations across the EU and the United States, available in public Google Earth satellite imagery. It includes 5,000 images containing a total of 45,096 objects across 15 different vehicle categories, including cars, trucks, buses, long vehicles, airplanes, ships, and other categories. For experimental purposes, we divided the images into training and validation sets in an 8:2 ratio, with 4,000 images in the training set and 1,000 images in the validation set, all randomly distributed. There is a significant class imbalance in the SIMD dataset, with smaller vehicles (such as cars and trucks) being more prevalent, while larger vehicles (such as buses, long vehicles, and airplanes) are relatively fewer. This imbalance may affect the detection performance of the model, so during the improvement of the YOLOv9 algorithm, data augmentation techniques were employed to enhance the model's performance in the face of class imbalance and multi-scale object detection, thereby improving the model's ability to recognize minority classes.

B. Experimental environment and parameter configuration

The experiments in this study were conducted on a server equipped with an NVIDIA GeForce RTX 3080Ti graphics card, which has 10GB of VRAM, effectively supporting the efficient training of deep learning models. The operating system was Windows 10, and the main software environment included CUDA 11.8, Python 3.8.10, and Pytorch 2.0.0. The model training was set for a total of 300 epochs. To prevent overfitting, the EarlyStopping strategy was employed with

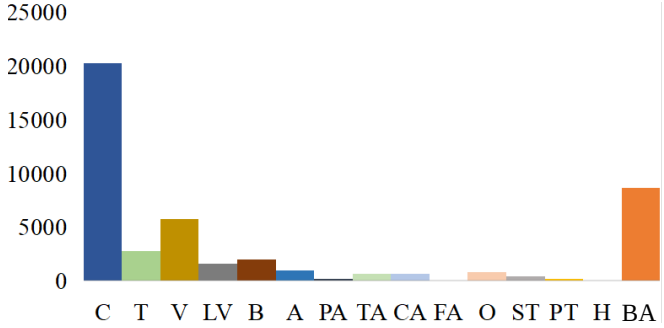

Fig. 4: The number of labels for each category in SIMD dataset.

a patience value of 50, meaning that if the validation loss did not show a significant decrease over 50 consecutive epochs, the training would be terminated early. The batch size was set to 4. Additionally, the learning rate was set to 0.01 to balance training speed and model convergence. Other parameters were kept at their default values.

C. Model evaluation metrics

To evaluate the performance of the YOLOv9-based object detection model on the SIMD dataset, this study employed several key metrics, including Precision, Recall, mean Average Precision (mAP), and F1-score. The definitions of these metrics are provided in Equation 11. Precision measures the proportion of correctly predicted positive instances, while Recall assesses the model's ability to detect actual positive samples. To balance Precision and Recall, the F1-score was used as a comprehensive metric. Additionally, mean Average Precision (mAP) was used to evaluate the overall performance of the model in multi-class detection tasks.

$$\begin{aligned} P &= \frac{TP}{TP+FP}, \\ R &= \frac{TP}{TP+FN}, \\ F1 &= 2 \times \frac{P \times R}{P+R}, \\ mAP &= \frac{1}{N}\sum_{i=1}^{N} AP_i. \end{aligned} \tag{10}$$

Through the calculation and analysis of these metrics, we can comprehensively evaluate the model's detection performance on the SIMD dataset and provide strong data support for the optimization and improvement of the model.

D. Results and Analysis of the Precision-Recall Curve

The Precision-Recall (P-R) curve in the figure illustrates the differences in detection performance of the YOLOv9 model on the SIMD dataset before and after improvements. The improved YOLOv9 model shows an increase in the mean Average Precision (mAP@0.5) across all categories from 0.864 to 0.880, with significant enhancements in both detection accuracy and recall. Specifically, the AP value for the 'pushbacktruck' category increased from 0.702 to 0.824, while the AP value for the 'propeller' category slightly decreased from 0.992 to 0.980. However, the AP values for most categories improved overall. Additionally, the improved model achieved a better balance between precision and recall across most categories, with a smoother curve shape, indicating more stable detection performance across different thresholds. Particularly for complex targets like the 'fighter' category, although the AP value remained at 0.995, the improved curve is more concentrated, reflecting the model's enhanced balance between precision and recall. Overall, the improved YOLOv9 model has effectively enhanced its object detection performance on the SIMD dataset.

E. Ablation experiments

To verify the impact of each module on the performance of our proposed improved YOLOv9 model on the SIMD dataset, we conducted ablation experiments, with the results shown in Table 1. In this ablation study, we tested the C3_CD_CGA module, DSConvRepNCSPELAN4 module, and CARAFE module separately. By progressively adding or removing these modules, we were able to observe their influence on the overall performance of the model. The experimental results in the table display the Precision, Recall, and mean Average Precision (mAP) for different combinations of modules.

The results show that the model performs best when all modules are fully retained, achieving a Precision of 88.5%, Recall of 86.3%, and mAP of 88.0%. This indicates that each module plays a crucial role in enhancing the model's performance. Further analysis reveals that when no modules are added, the model's mAP drops from 88.0% to 86.4%, using the YOLOv9 model as the baseline, demonstrating that these modules significantly contribute to improving detection accuracy. When only the C3_CD_CGA module is added, the mAP increases slightly to 87.2%, indicating that the C3_CD_CGA module provides some performance enhancement. Adding only the DSConvRepNCSPELAN4 module results in an mAP of 87.4% and a Recall of 86.8%, highlighting the module's important role in improving Recall. When both the C3_CD_CGA and DSConvRepNCSPELAN4 modules are retained, the model's mAP reaches 87.6%, showing a strong synergistic effect between these two modules. Adding only the CARAFE module results in an mAP of 87.2%, but when the CARAFE module is added to the model already containing the C3_CD_CGA and DSConvRepNCSPELAN4 modules, the mAP further improves to 88%, demonstrating CARAFE's significant contribution to enhancing accuracy. Overall, this ablation study clearly illustrates the performance improvements contributed by each module, validates the effectiveness of our proposed improved model, and provides a reference for further optimization research.

F. Comparison results of different models

To comprehensively evaluate the performance of our proposed object detection model for remote sensing images based on YOLOv9, we conducted detailed experiments on the SIMD dataset. Table 2 summarizes the experimental results of various models, providing a systematic comparison of Precision, Recall, Image Size, and mAP. The results show that our model performs excellently across all metrics, particularly achieving an mAP of 88.0%, significantly surpassing the baseline YOLOv9 model (86.4%). Compared to other models such as YOLO-DA, YOLO-SE, and MHLDeT, our

TABLE I: Ablation experiments

	C3_CD_CGA	DSConvRepNCSPELAN4	CARAFE	Precision/%	Recall/%	mAP
YOLOv9	-	-	-	87.8	85.6	86.4
YOLOv9	√	-	-	87.3	86.3	87.2
YOLOv9	-	√	-	87.2	86.8	87.4
YOLOv9	-	-	√	86.9	85.8	87.2
YOLOv9	-	√	√	87.2	87.4	87.6
YOLOv9	√	-	√	87.4	87.5	87.7
YOLOv9	√	√	-	88.1	86.5	87.6
YOLOv9	√	√	√	88.5	86.3	88.0

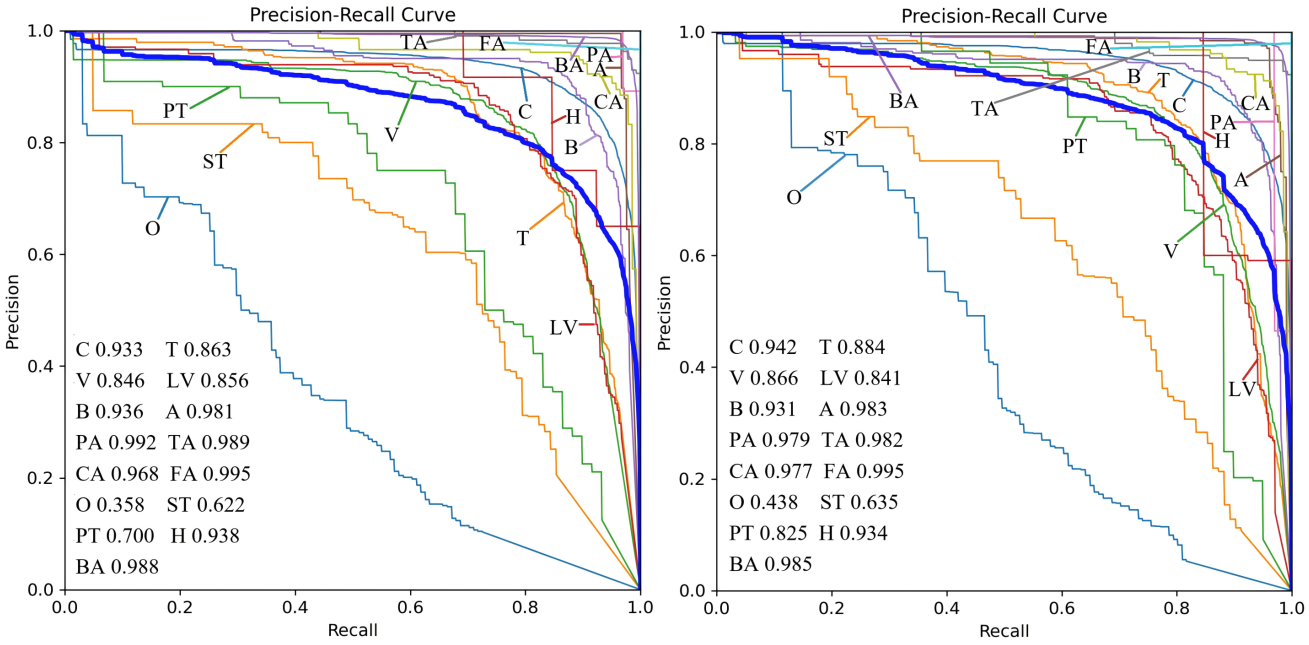

Fig. 5: Results Comparison

model achieved a Precision of 88.5% and a Recall of 86.3%, reflecting not only its advantage in recognition accuracy but also its strong capability in comprehensive object detection. In contrast, although YOLOv8, YOLOv10 and YOLOv11 offer faster computational speeds, their mAPs are only 79.3%, 81.1%and 81.0%, respectively, which are significantly lower than our model, indicating that they may fall short in high-precision tasks. Overall, our improved model demonstrates significant accuracy improvements on the SIMD dataset, laying a solid foundation for future research and applications.

TABLE II: Compare different categories pairwise

Method	ImageSize	Precision	Recall	mAP
YOLO-DA	640*640	-	-	80.6
YOLO-SE	640*640	83.6	81.9	86.5
MHLDeT	640*640	77.1	82.5	84.7
YOLOv8	640*640	71.5	77.5	79.3
YOLOv10	640*640	83.7	79.6	81.1
YOLOv11	640*640	81.1	78.9	81.0
ours	640*640	88.5	86.3	88.0

G. Random Image Detection

In remote sensing image datasets, the detection task is challenging due to typically small target objects and diverse categories. From the comparison images, it is evident that the improved model achieves more precise bounding box localization in complex backgrounds and dense target scenarios. Specifically, the original YOLOv9 model exhibited missed detections across all images and had noticeable false detections in the three image. These results indicate that the improved YOLOv9 model shows significant performance enhancement on the SIMD dataset, effectively reducing both false detections and missed detections. Overall, the detection accuracy is significantly improved, validating the effectiveness of our model improvements.

V. Conclusion

This paper proposes an improved algorithm based on YOLOv9, optimized for object detection in remote sensing images. By incorporating the C3_CD_CGA module, DSConvRepNCSPELAN4 module, and CARAFE module, we successfully developed a detection model that is better suited for aerial imagery on the foundation of the YOLOv9 algorithm. These modules integrate advanced feature extraction, feature grouping, multi-level feature fusion, and

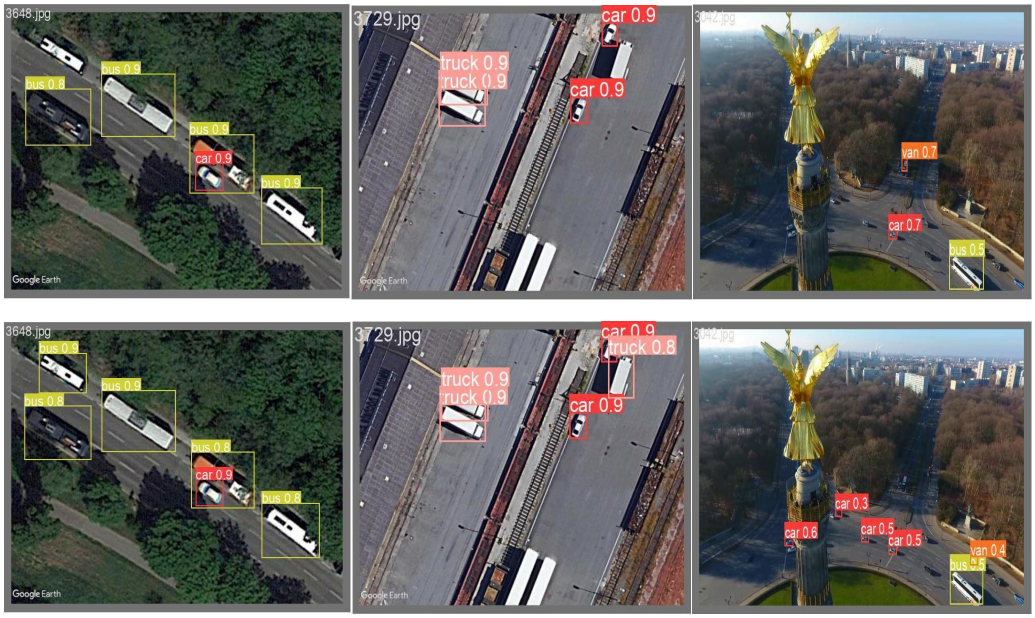

Fig. 6: Results Comparison

contextual information processing techniques, significantly enhancing the model's performance in complex backgrounds and multi-scale target detection. Experimental results on the SIMD dataset demonstrate a notable improvement in detection accuracy. The introduction of the C3_CD_CGA module enables the model to better capture both local and global features of the targets, achieving multi-class object detection. Meanwhile, the DSConvRepNCSPELAN4 and CARAFE modules further improve detection accuracy by optimizing the convolution and upsampling processes. Experimental results indicate that the improved YOLOv9 model achieves a 1.6% increase in mean accuracy on the SIMD dataset compared to the baseline YOLOv9 model. However, the improved YOLOv9 still faces some false positive issues when dealing with shadow occlusion. Overall, this study provides new ideas and technical foundations for more efficient object detection models in aerial imagery. Future work could explore solutions to address shadow occlusion issues and further achieve model lightweighting while maintaining detection accuracy.

REFERENCES

[1] J. Lin, Y. Zhao, S. Wang, *et al.*, "Yolo-da: An efficient yolo-based detector for remote sensing object detection," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[2] T. Wu and Y. Dong, "Yolo-se: Improved yolov8 for remote sensing object detection and recognition," *Applied Sciences*, vol. 13, no. 24, p. 12977, 2023.

[3] Y. Zhang, M. Ye, G. Zhu, *et al.*, "Ffca-yolo for small object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[4] X. Fan, Z. Hu, Y. Zhao, J. Chen, T. Wei, and Z. Huang, "A small ship object detection method for satellite remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[5] W. Teng, H. Zhang, and Y. Zhang, "X-ray security inspection prohibited items detection model based on improved yolov7-tiny," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 7, pp. 1279–1287, 2024.

[6] J. Pan and Y. Zhang, "Small object detection in aerial drone imagery based on yolov8.," *IAENG International Journal of Computer Science*, vol. 51, no. 9, pp. 1346–1354, 2024.

[7] Y. Zhang, H. Zhang, Q. Huang, Y. Han, and M. Zhao, "Dsp-yolo: An anchor-free network with dspan for small object detection of multiscale defects," *Expert Systems with Applications*, vol. 241, p. 122669, 2024.

[8] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[9] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13708–13715, IEEE, 2021.

[10] S. Shen, Z. Liu, B. Zhao, *et al.*, "Improving real-world object detection using balanced loss," in *2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–5, IEEE, 2020.

[11] R. Xu, C. Chen, J. Peng, *et al.*, "Toward raw object detection: A new benchmark and a new model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13384–13393, 2023.

[12] C. Xu, T. Zhang, D. Zhang, *et al.*, "Deep generative adversarial reinforcement learning for semi-supervised segmentation of low-contrast and small objects in medical images," *IEEE Transactions on Medical Imaging*, 2024.

[13] B. Liu *et al.*, "Small object detection using multi-scale feature fusion and attention," in *2022 41st Chinese Control Conference (CCC)*, IEEE, 2022.

[14] T. Chen, D. Wang, W. Tao, *et al.*, "Cassod-net: Cascaded and separable structures of dilated convolution for embedded vision systems and applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3182–3190, 2021.

[15] X. Liu, H. Peng, N. Zheng, *et al.*, "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14430, 2023.

[16] Y. Qi, Y. He, X. Qi, *et al.*, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6070–6079, 2023.

[17] J. Wang, K. Chen, R. Xu, *et al.*, "Carafe: Content-aware reassembly of features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3007–3016, 2019.

[18] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, *et al.*, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 390–391, 2020.

[19] S. Liu, L. Qi, H. Qin, *et al.*, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768, 2018.

[20] T.-Y. Lin, P. Dollár, R. Girshick, *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.