A Lightweight Object Detection Algorithm for Remote Sensing Images

Donghao Hou, Yujun Zhang*, Jia Ren

***Abstract*—With the continuous advancement of satellite technology, remote sensing images has been increasingly applied in fields such as urban planning, environmental monitoring, and disaster response. However, remote sensing images often feature small target sizes and complex backgrounds, posing significant computational challenges for object detection tasks. To address this issue, this paper proposes a lightweight remote sensing images object detection algorithm based on YOLOv9. The proposed algorithm incorporates the SimRMB module, which effectively reduces computational complexity while improving the efficiency and accuracy of feature extraction. Through a dynamic attention mechanism, SimRMB is capable of focusing on important regions while minimizing background interference, and by integrating residual learning and skip connections, it ensures the stability of deep networks. To further enhance detection performance, the FasterRepNCSPELAN4 module is introduced, which employs PConv operations to reduce computational load and memory usage. It also utilizes dilated convolutions and DFC attention mechanisms to strengthen feature extraction, thereby increasing the efficiency and accuracy of object detection. Additionally, this study integrates the GhostModuleV2 module, which generates core feature maps and employs lightweight operations to create redundant features, greatly reducing the computational complexity of convolutions.Experimental results show that on the SIMD dataset, the improved YOLOv9 model has a parameter size of 167.88 MB and GFLOPs of 208.6. Compared to the baseline YOLOv9 model (parameter size: 194.57 MB, GFLOPs: 239.0), the parameter size is reduced by 13.71%, GFLOPs are reduced by 12.72%, and detection accuracy is improved by 1.4%. These results demonstrate that the proposed lightweight YOLOv9 model effectively reduces computational overhead while maintaining excellent detection performance, providing an efficient solution for object detection tasks in resource-constrained environments.**

***Index Terms*—Attention mechanism, Lightweight, Object detection, YOLOv9.**

I. Introduction

With the rapid advancement of satellite technology, remote sensing images has become indispensable in areas such as urban planning, environmental monitoring, and disaster response. UAVs provide great convenience across industries with their flexibility, multi-angle perspectives, and efficient data acquisition capabilities. However, small object detection in remote sensing images still faces significant challenges [1, 2], especially in resource-constrained environments where computational power and battery capacity are limited, such as on UAVs. The task of detecting small objects in aerial images is further complicated by high resolution, high object density, and complex backgrounds, which often impose high computational demands.

Traditional large-scale network models can offer good detection accuracy, but their complex architectures and large parameter sizes lead to slower inference speeds, consuming a substantial amount of computational resources and energy, making them unsuitable for scenarios that require real-time performance and portability. In recent years, with the rapid development of deep learning theories and techniques, deep learning-based object detection algorithms have significantly outperformed traditional methods in general detection tasks. Deep learning-based object detection algorithms are mainly divided into two categories: two-stage detectors (such as the RCNN series) and one-stage detectors (such as the YOLO series and SSD) [3–5]. One-stage detectors excel in computational efficiency but typically struggle with small object detection and localization; on the other hand, two-stage detectors achieve higher detection accuracy by first locating and then recognizing objects, but their real-time performance is lacking. Therefore, achieving model lightweighting while improving the performance of small object detection in remote sensing images under complex backgrounds has become a key research focus.

Tan proposed optimizing model depth, width, and resolution simultaneously through a compound scaling strategy, which further reduced computational complexity. The introduction of the Fused-MBConv structure significantly accelerated training speed on the basis of a lightweight model, making it particularly suitable for tasks requiring fast training and efficient inference, though accuracy did not improve substantially [6]. Tan also further improved the EfficientNet series by re-optimizing the compound scaling strategy, enhancing performance across various computational resource conditions. New activation functions and improved feature fusion methods were introduced, enabling EfficientNetV3 to maintain classification and detection accuracy while reducing computational load and parameter size, making it ideal for devices with limited computational resources [7]. Ding proposed a new re-parameterization structure that uses a multi-branch architecture during training to enhance network accuracy, while converting it to a single-branch convolution during inference, significantly improving inference efficiency. Through this approach, RepVGG achieved a balance between efficient inference and superior performance while maintaining a VGG-like architecture [8]. Ma introduced channel grouping convolution and channel shuffle operations, which significantly improved inference speed by reducing computational load and memory access

Manuscript received October 30, 2024; revised January 25, 2024. 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).

Jia Ren is a graduate student of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: 18341275676@163.com).

[9]. Yu combined the advantages of Transformer structures and optimized the attention mechanism through lightweight design while using convolutional feature extraction to reduce parameter size. LightViT retains the strong global receptive field and self-attention mechanism of Transformers, while structural adjustments significantly reduce the model's computational complexity and memory usage, making it suitable for scenarios requiring efficient inference [10]. Liu proposed a lightweight image segmentation model for mobile devices, combining MobileNet's efficient convolutions and the SAM (Segment Anything Model) framework for efficient segmentation tasks. MobileSAM, through the introduction of depthwise separable convolutions and lightweight feature extraction modules, significantly reduced computational complexity while maintaining efficient inference and high-quality segmentation performance in resource-constrained environments [11]. Although these algorithms achieve lightweight object detection in remote sensing images applications, they still face significant challenges.

To address these issues, this paper proposes a lightweight remote sensing images object detection algorithm based on YOLOv9 [12], aiming to improve detection efficiency and accuracy under complex backgrounds and resource-constrained environments by optimizing network architecture and feature extraction modules. Specifically, a SimRMB module is proposed to replace the RepNCSPELAN4 module in the YOLOv9 backbone network [13, 14]. The lightweight design of the SimRMB module not only significantly reduces computational complexity but also, through the introduction of a dynamic attention mechanism, effectively focuses on key areas in images, reduces background interference, and improves detection accuracy. By integrating residual learning and skip connections, the network stability is enhanced, along with improved feature extraction efficiency. In the head network, the RepNCSPELAN4 module is replaced by the FasterRepNCSPELAN4 module [15], which combines the advantages of PConv and RepNCSPELAN4, achieving more efficient feature extraction and inference speed, reducing computational load and memory usage without sacrificing feature quality. The GhostModuleV2 module is also introduced [16], which generates redundant features through inexpensive operations to simulate a complete convolutional output, significantly reducing computational complexity and memory consumption. Through the combined optimization of these modules, the proposed method significantly reduces the computational complexity and memory usage of the model while enhancing detection accuracy and inference capabilities.

II. Related work

Remote sensing images object detection is one of the crucial research directions in the current field of computer vision, especially in applications such as urban planning, agricultural monitoring, and disaster response. However, remote sensing images often feature dense small objects, significant scale variations, and complex backgrounds, which pose substantial challenges for traditional detection algorithms. To enhance detection performance and computational efficiency, researchers have gradually introduced lightweight models and attention mechanisms into UAV and aerial imaging detection tasks [17–19].

In recent years, the YOLO series models (such as YOLOv4, YOLOv7, and YOLOv8) have made remarkable progress in single-stage detection frameworks, achieving high detection speed and accuracy [20, 21]. Nevertheless, they still have certain limitations in handling small object detection. To further reduce computational complexity and adapt to resource-constrained environments like UAV aerial imagery, various lightweight models have been proposed. For instance, GhostNet generates redundant features through inexpensive operations, thereby reducing convolutional overhead while improving inference speed [22]. Additionally, the MobileNet series, through depthwise separable convolutions, significantly reduces computational cost and performs exceptionally well on mobile and embedded devices.

In the domain of object detection, EfficientNet leverages a compound scaling strategy to optimize model depth, width, and resolution, achieving a balance between computational efficiency and accuracy [23]. Meanwhile, models like SlimYOLOv7 further reduce the number of convolutional layers and parameters, improving real-time detection performance, making it particularly suitable for small object detection in UAV aerial imagery [24]. Models such as NanoDet and MobileSAM have also proposed specific optimization strategies targeting small object detection and resource-constrained scenarios, effectively improving detection efficiency.

Moreover, recent studies have highlighted the importance of attention mechanisms in aerial image detection. Models like MobileViT, which combine Transformer structures with convolutional neural networks, enhance global feature capturing capabilities while effectively reducing computational complexity, making them suitable for real-time detection tasks.

Building on these research outcomes, this study proposes a lightweight aerial image object detection algorithm based on YOLOv9, which incorporates the latest lightweight design and attention mechanisms. This approach significantly reduces computational resource consumption while maintaining high detection accuracy, providing an efficient solution for small object detection in complex scenes.

III. Method Introduction

A. Modules of the Improved YOLOv9 Algorithm

To address these challenges, we not only applied lightweight strategies but also achieved significant accuracy improvement through the proposed SimRMB module and the introduction of the FasterRepNCSPELAN4 and GhostModuleV2 modules.

In the backbone network, the RepNCSPELAN4 module was replaced with the SimRMB module. The SimRMB module reduces unnecessary convolutional layers and parameters, effectively lowering computational costs and improving inference speed, while optimizing scale invariance of target detection. It focuses on important regions during feature extraction, ignoring irrelevant background noise. By adopting a multi-branch convolutional design, different branches process features at different scales in parallel, enhancing the model's capability to capture small objects and improving detection accuracy in complex scenes. In the neck network, the FasterRepNCSPELAN4 and GhostModuleV2 modules were introduced. The FasterRepNCSPELAN4 module, by combining PConv and re-parameterization structures,

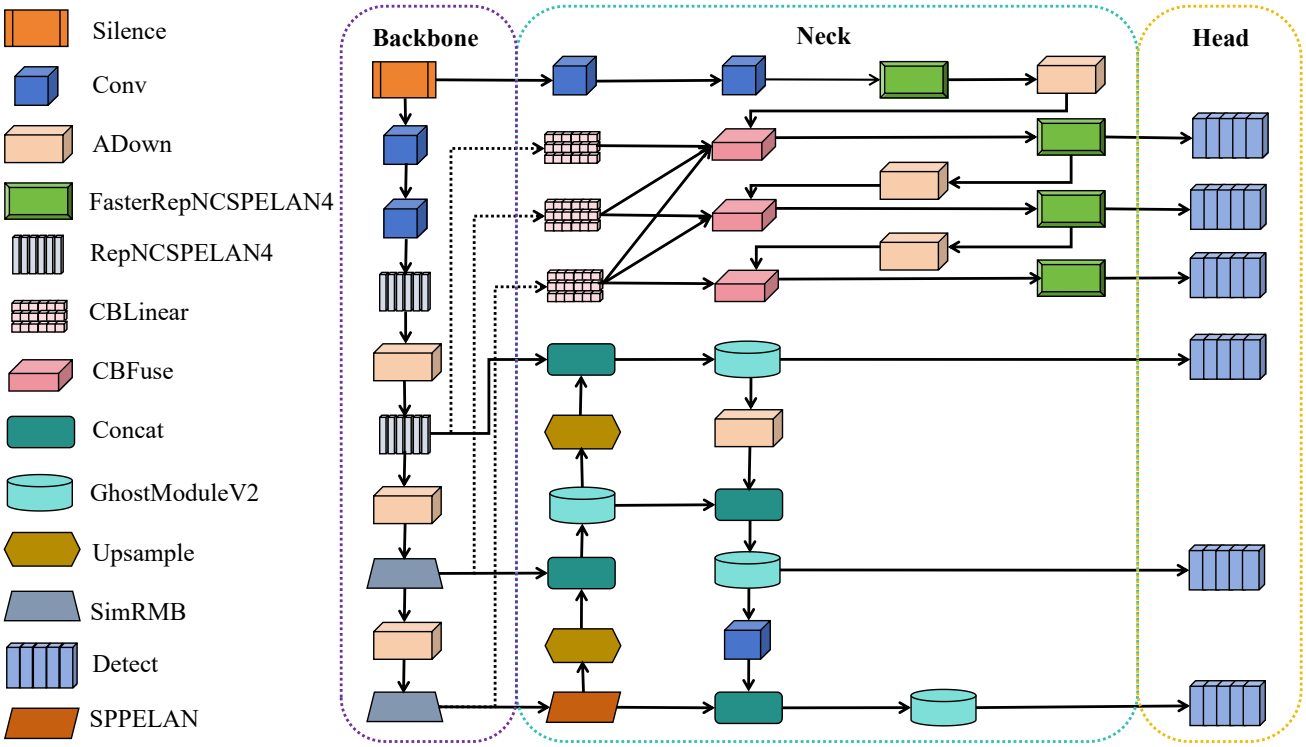

Fig. 1: Overall improved architecture diagram

achieved efficient feature extraction and optimized inference. It uses partial convolutions to reduce unnecessary convolutional computations and memory access, significantly lowering computational costs and parameters while maintaining model accuracy. The GhostModuleV2 module generates core feature maps through a few standard convolution operations, followed by a series of inexpensive operations (such as depthwise convolutions) to produce redundant features, thereby simulating the output of standard convolutions. This design significantly reduces both computational load and memory usage.

B. SimRMB

Due to the complexity of the environment in the SIMD remote sensing images dataset, along with diverse categories and varied information scales, high computational resources are required. The SimRMB module simplifies the residual structure, reducing unnecessary convolutional operations and feature map size changes, thereby accelerating the inference process. This makes it particularly suitable for high-resolution aerial image processing. Furthermore, SimRMB adopts a combination of depthwise and pointwise convolutions, effectively reducing the floating-point operations (FLOPs) while maintaining strong feature extraction capabilities, significantly decreasing the computational burden.

The SimRMB module is an efficient and lightweight module designed to enhance feature extraction and object detection capabilities in dense prediction tasks. It combines the characteristics of skip connections, residual learning, and both local and global feature modeling, specifically optimizing computational complexity to facilitate inference on resource-constrained devices. SimRMB applies dynamic attention mechanisms to weight features, significantly reducing computational cost while maintaining performance. The core idea of the SimRMB module lies in the combination of feature reuse and lightweight convolutions. First, the input features undergo a normalization operation, followed by a series of lightweight convolutions to extract local features. Let the input feature be $X \in R^{B\times C\times H\times W}$, where B is the batch size, C is the number of channels, and H and W represent the height and width of the feature map, respectively. The normalization process applied to the input is shown in Equation (1).

$$X_{\text{norm}} = \text{Norm}(X) \tag{1}$$

SimRMB introduces a dynamic attention mechanism to enhance feature representation capability. For each element in the feature map, SimRMB calculates its relative importance to the surrounding neurons. An energy function, E(x) is defined to measure the differences between each position in the feature map and its surrounding positions, with the formulation given in Equation (2). This weighting method, based on the energy function, allows the model to dynamically assign an attention value to each pixel in the feature map, ensuring that the model accurately focuses on important regions while ignoring irrelevant background information.Here, X_i represents the current neuron in the feature map, μ is the mean of all neurons, σ^2 is the variance of the feature map, λ is a hyperparameter used to adjust the flexibility of the energy function, and n is the total number of neurons in the feature map.

$$E(x_i) = \frac{1}{n}\sum_{j\neq i}\left(\frac{(x_j-\mu)^2}{\sigma^2+\lambda}\right) + 0.5 \tag{2}$$

The attention-weighted feature map is further processed through depthwise convolution to extract local features, with

the depthwise convolution operation defined in Equation (3). Here, W represents the convolution kernel, and X_{att} is the attention-weighted feature map. Depthwise convolution effectively captures local detail information while maintaining lightweight computational complexity.

$$X_{\text{conv}} = \text{Conv}(X_{\text{att}}) = W * X_{\text{att}} \tag{3}$$

SimRMB also employs skip connections and residual learning strategies, where the residual connection is formed by adding the input X to the output feature X_{conv}, as shown in Equation (4). Here, DropPath is a regularization strategy used to randomly drop certain paths during training, thereby enhancing the model's generalization capability. This residual connection design facilitates effective information RMBflow and helps prevent the vanishing gradient problem, ensuring that the SimRMB module maintains stable performance even in deep networks.

$$X_{\text{out}} = X + \text{DropPath}(X_{\text{conv}}) \tag{4}$$

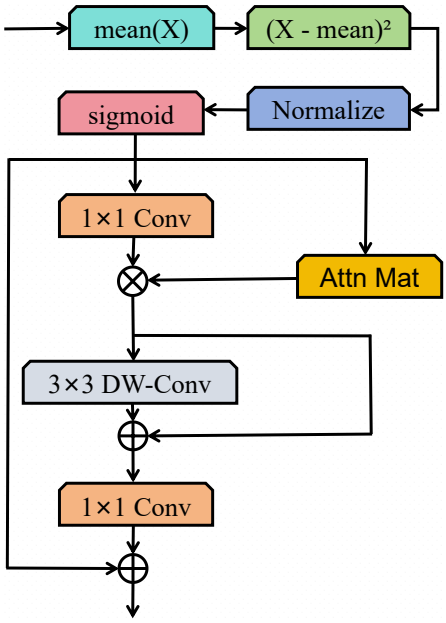

Fig. 2: SimRMB network architecture

C. FasterRepNCSPELAN4

The RepNCSPELAN4 module uses depthwise separable convolutions, feature map fusion, and multi-scale feature extraction to accurately capture different targets in complex scenes. However, when dealing with large-scale data, neural networks still face bottlenecks in computational performance and inference speed. To address this issue, the FasterRepNCSPELAN4 module integrates the partial convolution mechanism (PConv) from FasterNet with the RepNCSPELAN4 module, significantly enhancing the inference speed and feature extraction capabilities of the YOLOv9 model. This module effectively reduces convolutional computations and memory access requirements through partial convolution, thereby greatly improving efficiency when processing large-scale aerial imagery data.The introduction of PConv not only reduces the computational complexity of standard convolution and decreases floating-point operations (FLOPs), but also generates additional features by integrating the lightweight Ghost module, significantly reducing the computational burden. Additionally, FasterRepNCSPELAN4 reduces intermediate layer normalization and redundant calculations, enabling YOLOv9 to achieve efficient inference and low-power operation even on resource-constrained devices, which greatly enhances model deployment efficiency. Lastly, the multi-branch feature fusion structure of FasterRepNCSPELAN4 incorporates multi-path information flow at different network levels, not only preserving and fusing more input features but also effectively avoiding common problems in deep networks, such as gradient vanishing or explosion. This architecture enables YOLOv9 to maintain high-precision object detection while further improving inference speed and reducing computational complexity. The FasterRepNCSPELAN4 module is shown in Figure 3.

The input feature $X \in R^{H\times W\times C}$ is first compressed through a 1×1 convolution, which aims primarily to reduce the number of channels in subsequent computations, thereby decreasing the overall computational load. The output is denoted as $X' \in R^{H\times W\times C'}$, where C' represents the compressed number of channels. The corresponding formula is shown in Equation (5).

$$X' = \text{Conv}_{1\times 1}(X) \tag{5}$$

The RepNCSP module combines the advantages of depthwise separable convolution and standard convolution, maintaining flexibility during training while reducing computational complexity during inference through parameter merging. This module effectively extracts multi-scale features while keeping computational overhead low. The Partial Convolution (PConv) is used to reduce feature computation complexity by applying convolution operations only to a subset of channels while keeping the rest unchanged. This not only reduces FLOPs but also avoids redundant computations, thereby accelerating inference speed, as shown in Equation (6). The final fused features are then compressed again in terms of channel count through a 1×1 convolution layer, as described in Equation (7).

$$\begin{cases} Y_1' = \text{PConv}(\text{RepNCSP}(X_1')) \\ Y_2' = \text{PConv}(\text{RepNCSP}(X_2')) \end{cases} \tag{6}$$

$$Z = \text{Conv}_{1\times 1}(\text{Concat}(Y_1', Y_2')) \tag{7}$$

The FasterRepNCSPELAN4 module, by integrating partial convolution (PConv), the RepNCSP module, and a multi-path feature fusion structure, significantly enhances YOLOv9 in terms of inference speed, computational efficiency, and feature extraction capabilities. It is not only well-suited for complex aerial imagery scenarios but also capable of efficient deployment on resource-constrained devices, achieving lightweight yet high-performance detection.

D. GhostModuleV2

Due to the characteristics of remote sensing images, including multiple classes, small target sizes, and complex backgrounds, higher computational resources are required. The model must maintain high accuracy while achieving

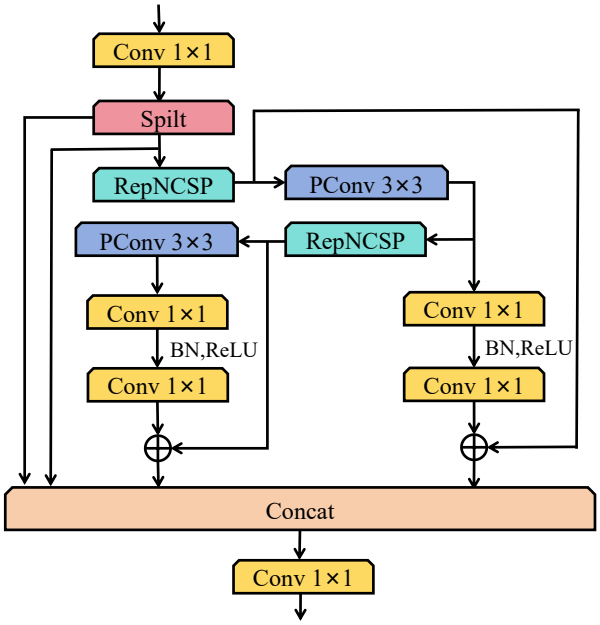

Fig. 3: FasterRepNCSPELAN4 network architecture

efficient inference speed and minimizing computational demands. To address these challenges, this study incorporates the GhostNetV2 module into the YOLOv9 architecture.

The DFC (Decoupled Fully Connected) attention mechanism in the GhostNetV2 module aggregates features in both horizontal and vertical directions, effectively capturing global spatial information and improving accuracy. It also avoids complex tensor transposition and reshaping operations, thereby increasing the actual inference speed, enhancing efficiency, and reducing computational complexity, achieving module lightweighting. Addresses deficiencies in capturing spatial information, allowing generated features to extend beyond single-channel convolution operations, thereby enhancing the global dependencies of features. By employing cost-effective operations such as depthwise convolution, it further reduces computational load and parameter count, making it optimized for memory-constrained environments and ensuring efficient inference. The DFC module and GhostNetV2 module introduced in this study is illustrated in Figure 4 and Figure 5.

Decoupling the traditional fully connected (FC) layer into horizontal and vertical FC layers allows the FC layer to generate an attention map, as shown in Equation (8). This captures long-range spatial dependencies in both directions, not only enhancing the ability to capture global information but also significantly reducing computational complexity.

$$\begin{aligned} a'_{hw} &= \sum_{h'} F^{H}_{h,h'} z_{h'w} \\ a_{hw} &= \sum_{w'} F^{W}_{w,w'} a'_{hw'} \end{aligned} \tag{8}$$

GhostNetV2 utilizes a 1×1 convolution to generate the initial features, followed by inexpensive depthwise convolutions to generate additional features, which are then concatenated. The formula is shown in Equation (9), where $*$ denotes the convolution operation, and $F_{1\times1}$ represents the pointwise convolution.

$$Y = \text{Concat}(X * F_{1\times1}, (X * F_{1\times1}) * F_{\text{dp}}) \tag{9}$$

In GhostNetV2, the DFC attention mechanism works in parallel with the first Ghost module, enhancing the representational capacity of the expanded features. The final output is the element-wise product of the outputs from the Ghost module and the DFC attention, as shown in Equation (10), where σ represents the Sigmoid function and $\odot$ indicates element-wise multiplication.

$$O = \sigma(A) \odot V(X) \tag{10}$$

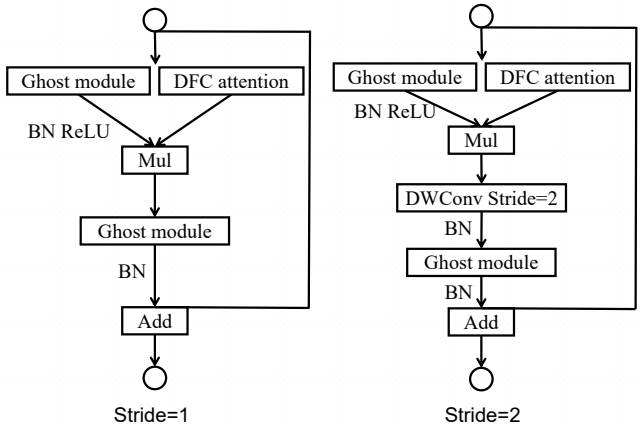

Fig. 4: GhostModuleV2 network architecture

IV. Experimental Design and Implementation

A. Experimental environment and parameter configuration

The experiments in this study were conducted on a server equipped with an NVIDIA GeForce RTX 3080Ti GPU, which provides excellent computational performance suitable for efficiently executing deep learning tasks. The experiment was configured with 300 training epochs, and an EarlyStopping strategy was introduced to effectively prevent model overfitting. If the validation loss did not show a significant decrease for 50 consecutive epochs, the training process was terminated early to improve efficiency and enhance model generalizability. During training, the batch size was set to 4, and the initial learning rate was set to 0.01 to balance training speed and model convergence. Except for the aforementioned configurations, all other hyperparameters were kept at their default settings. The hardware and software environments and configuration parameters used in the experiments are detailed in Table 1.

TABLE I: Experimental environment

Environment	Configuration
Operation platform	Windows 10
GPU	NVIDIA GeForce RTX 3080Ti
Programming language	Python 3.8.10
Deep learning framework	CUDA 11.8,Pytorch 2.0.0
Video memory	10G

B. Dataset Introduction

All experiments in this study were conducted on the SIMD dataset to ensure fair performance comparison under

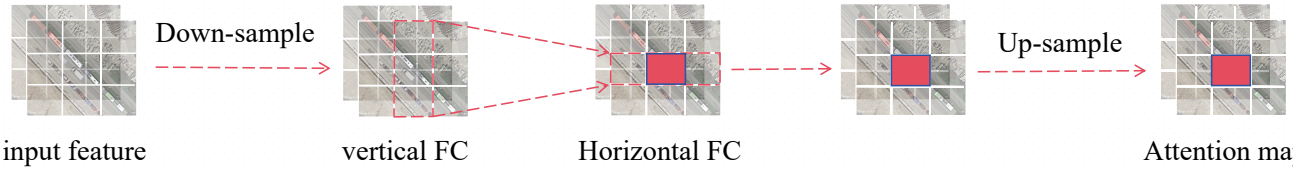

Fig. 5: The information flflow of DFC attention

identical conditions. The images in the dataset were obtained from public Google Earth satellite imagery, covering various geographical locations across Europe and the United States. The dataset contains 5000 images, of which 4000 are used for the training set and 1000 for the validation set. A total of 45,096 target objects were annotated, covering 15 different vehicle categories, including cars, trucks, buses, long vehicles, as well as various types of airplanes and ships. These rich annotations make the SIMD dataset valuable for tasks such as vehicle detection and classification.

C. Comparison results of different models

To evaluate the performance of the improved YOLOv9 model, this study conducted comparative experiments on the SIMD dataset against several mainstream lightweight object detection models, including MHLDDet, YOLOv7-Tiny, YOLO-SE, YOLO-DA,and YOLOv9. The experimental results are shown in Table 3.

Compared to other models, the improved YOLOv9 model demonstrated superior performance in all metrics. Specifically, with the same input image size (640×640), the improved YOLOv9 model achieved a mean Average Precision (mAP) of 87.8%, showing a significant improvement over the original YOLOv9's 86.4%, and outperforming other models such as YOLOv7-Tiny, YOLO-SE, and YOLO-DA. Meanwhile, the number of parameters was reduced to 44.03M, representing a 13.71% reduction compared to the YOLOv9 baseline, and the computational cost (GFLOPs) was reduced by 12.72%, further validating the lightweight nature of the model. YOLOv7-Tiny, YOLOv11, and MHLDDet have smaller parameter counts, but their detection accuracies (mAP) are 82.3%, 81.0%, and 84.7%, respectively, which are slightly lower than the improved YOLOv9. YOLO-SE and YOLO-DA, although having larger parameter counts, still did not surpass the improved YOLOv9 in accuracy.

In summary, the improved YOLOv9 (Our) model achieved a good balance in terms of mAP, parameter count, and GFLOPs, providing high detection accuracy while maintaining low computational cost, thereby demonstrating strong lightweight and efficient characteristics.

TABLE II: Compare different categories pairwise

Method	ImageSize	Params(M)	GFLOPs	mAP
YOLO-DA	640*640	94.4	-	80.6
YOLO-SE	640*640	13.9	-	70.7
YOLOv7-Tiny	640*640	6.05	13.3	82.3
YOLOv9	640*640	51.03	239	86.4
YOLOv11	640*640	9.89	6.5	81.0
MHLDeT	640*640	5.28	12.2	84.7
ours	640*640	44.03	208.6	87.8

D. Ablation experiments

After introducing the FasterRepNCSPELAN4 module to the baseline model, the mean average precision (mAP) increased to 86.8%, while the number of parameters and GFLOPs were reduced to 48.05M and 225.6, respectively. This module significantly reduced model complexity while improving feature extraction efficiency.The melting experiment is shown in Table 2.

After introducing the FasterRepNCSPELAN4 module to the baseline model, the mean average precision (mAP) increased to 86.8%, while the number of parameters and GFLOPs were reduced to 48.05M and 225.6, respectively. This module significantly reduced model complexity while improving feature extraction efficiency.

When only the SimRMB module was added to the baseline model, mAP increased to 87.5%, while the number of parameters and GFLOPs decreased to 47.44M and 232.2. When the SimRMB module was added to the model that already contained FasterRepNCSPELAN4 and GhostModuleV2, mAP significantly increased to 87.8%, and the parameters and GFLOPs decreased to 44.03M and 208.6, respectively, demonstrating that the attention mechanism in SimRMB effectively captures key features, further improving detection accuracy and reducing computational cost.

In summary, the synergistic effects of the FasterRepNCSPELAN4, GhostModuleV2, and SimRMB modules significantly improved model detection accuracy while greatly reducing parameters and computational cost, thereby achieving lightweight and efficient object detection.

E. Random Image Detection

To comprehensively evaluate the performance of the improved YOLOv9 in practical applications, we randomly selected several images from the validation set for detection experiments and compared the results with those of the YOLOv9 baseline model, as shown in Figure 5. The images on the left show the detection results of the YOLOv9 baseline model, while the images on the right present the results of the improved YOLOv9. By comparison, it can be observed that the improved YOLOv9 demonstrates significant enhancements in detection accuracy and detail feature capture, particularly in detecting small objects in complex scenes, achieving higher accuracy. Some small or partially occluded targets are difficult to detect accurately, whereas the improved YOLOv9 effectively enhances feature extraction and attention allocation, enabling the model to capture target details and edge information more precisely. Additionally, the detection results of the improved YOLOv9 are more refined and stable, especially in scenarios where the background and target object textures are similar, resulting in significantly reduced false positives and missed detections compared to the original version.

TABLE III: Ablation experiments

	FasterRepNCSPELAN4	GhostModuleV2	SimRMB	Precision/%	Recall/%	mAP	Params(M)	GFLOPs
YOLOv9	-	-	-	87.8	85.6	86.4	51.03	239
YOLOv9	√	-	-	88.5	85.5	86.8	48.05	255.6
YOLOv9	-	√	-	88	86.9	86.9	50.61	229
YOLOv9	-	-	√	85.1	88.5	87.5	47.44	232.2
YOLOv9	√	√	-	88.6	85.5	87.2	47.63	215.5
YOLOv9	√	√	√	85.5	87.6	87.8	44.03	208.6

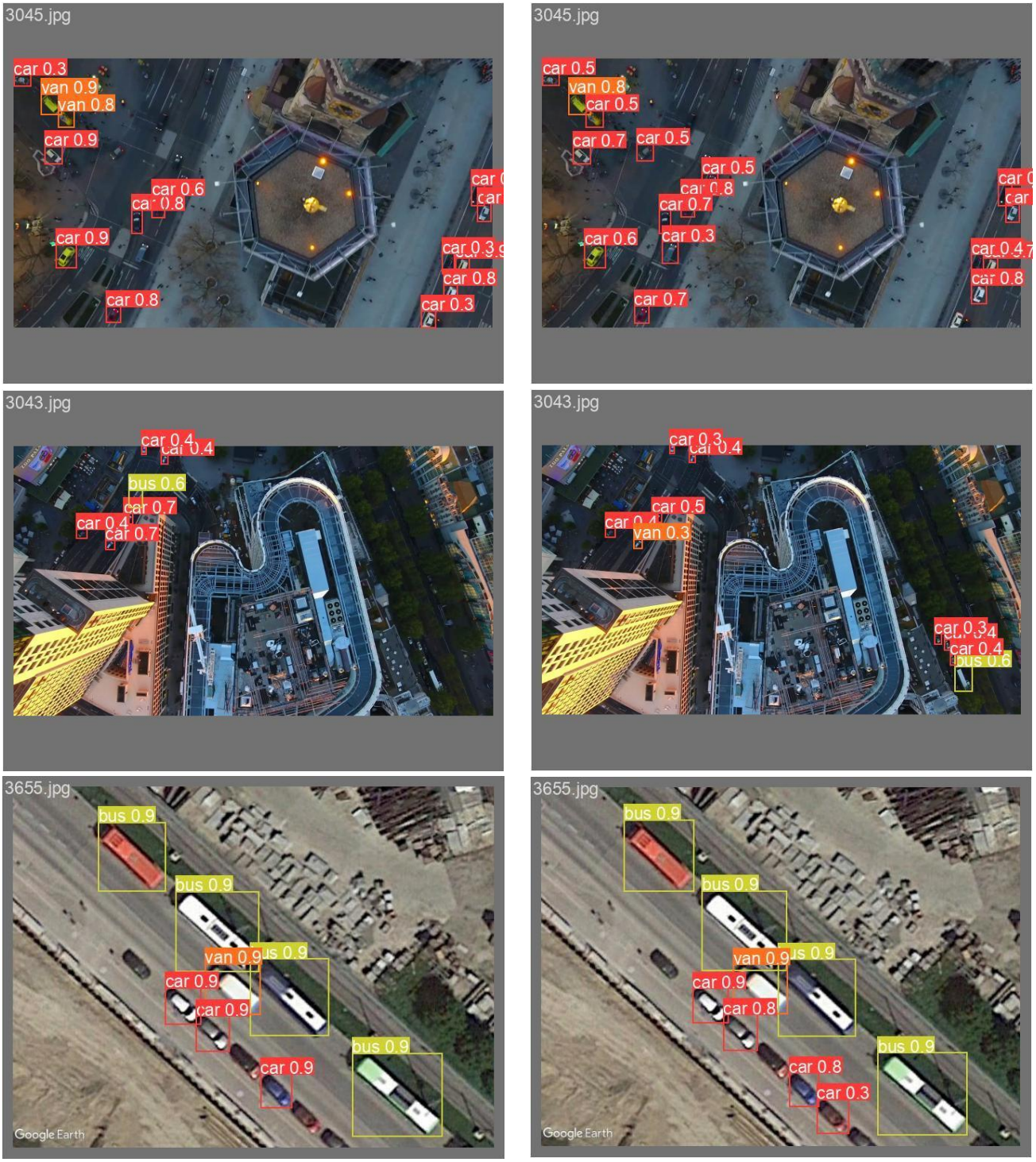

Fig. 6: Results Comparison

F. Model evaluation metrics

To comprehensively evaluate the performance of different model algorithms, this study adopts several metrics, including model size (in MB) to assess storage requirements, Number of Parameters to evaluate model complexity, and Floating Point Operations (FLOPs) to quantify computational cost. Additionally, the mean Average Precision (mAP) is used to assess the overall performance of the model in object detection tasks, ensuring that the model maintains high detection accuracy while achieving lightweight design.

V. Conclusion

In summary, the experimental results demonstrate that the improved YOLOv9 model achieves high-precision detection of remote sensing images targets while maintaining a lightweight design. By introducing the SimRMB module into the backbone network, the model enhances feature attention allocation and extracts key features through efficient convolution operations, thereby significantly reducing computational burden while improving detection accuracy. The FasterRepNCSPELAN4 module integrates the PConv from FasterNet with YOLOv9's RepNCSPELAN4, enabling efficient feature extraction and mixing, reducing computational complexity, and enhancing feature capture capability in complex scenes. GhostModuleV2 incorporates DFC attention to effectively capture long-range dependencies, further enhancing global feature representation. Compared to the baseline model, the improved YOLOv9 model achieved a 1.4% increase in mean Average Precision (mAP), a 7M reduction in the number of parameters, and a 30.4 reduction in GFLOPs, thus achieving a favorable balance between lightweight design and detection accuracy. Future work will involve further testing of the improved model on larger and more complex remote sensing datasets to validate its robustness and generalization capabilities across various application scenarios.

References

[1] 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.

[2] 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.

[3] X. Zhang and Z. Zhang, "Traffic sign detection algorithm based on improved yolov7," in *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, vol. 12707, pp. 1258–1266, SPIE, 2023.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37, Springer, 2016.

[5] Z. Wang, J. Zhang, Z. Zhao, and F. Su, "Efficient yolo: A lightweight model for embedded deep learning object detection," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2020.

[6] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, pp. 10096–10106, PMLR, 2021.

[7] M. Tan and Q. V. Le, "Efficientnetv3: Improved architectures and efficient training," *arXiv preprint arXiv:2303.07268*, 2023.

[8] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13733–13742, 2021.

[9] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

[10] L. Yu, M. Liu, J. Zhang, and J. Wang, "Lightvit: A lightweight vision transformer with efficient attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] Y. Liu, Y. Zhao, X. Zhang, X. Wang, C. Lian, J. Li, P. Shan, C. Fu, X. Lyu, L. Li, *et al.*, "Mobilesam-track: Lightweight one-shot tracking and segmentation of small objects on edge devices," *Remote Sensing*, vol. 15, no. 24, p. 5665, 2023.

[12] 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.

[13] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*, pp. 11863–11874, PMLR, 2021.

[14] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, Y. Wang, and C. Wang, "Rethinking mobile block for efficient attention-based models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1389–1400, IEEE Computer Society, 2023.

[15] J. Chen, S.-h. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan, "Run, don't walk: chasing higher flops for faster neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12021–12031, 2023.

[16] Y. Tang, K. Han, J. Guo, C. Xu, C. Xu, and Y. Wang, "Ghostnetv2: Enhance cheap operation with long-range attention," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9969–9982, 2022.

[17] 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.

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

[19] X. Wen, Y. Yao, Y. Cai, Z. Zhao, T. Chen, Z. Zeng, Z. Tang, and F. Gao, "A lightweight st-yolo based model for detection of tea bud in unstructured natural environments," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 3, pp. 342–349, 2024.

[20] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464–7475, 2023.

[21] S. Xie, M. Zhou, C. Wang, and S. Huang, "Csppartial-yolo: A lightweight yolo-based method for typical objects detection in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[22] K. Han, Y. Wang, Q. Tian, *et al.*, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[24] P. Zhang, Y. Zhong, and X. Li, "Slimyolov3: Narrower, faster and better for real-time uav applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.