Real-Time High-Precision Detection Technology for Aircraft Target in SAR Image Based on YOLOv9 and YOLOv10

Zhitan Zhou, Qichen Zheng, Qiang Yang(✉), and Yitao Ma

Harbin Institute of Technology, Harbin, China
yq@hit.edu.cn

Abstract. With the rapid development and application of SAR (Synthetic Aperture Radar) imaging technology, SAR-related technologies are also evolving rapidly. Among them, SAR image target detection, as an important branch, has shown significant application value. However, due to issues such as the high presence of speckle noise in SAR images, target blurring caused by target movement, and image distortion resulting from elevation differences, SAR target detection poses great difficulties, even for human visual inspection. Therefore, there is a need for an accurate and rapid detection method to complete target detection in SAR images. Since the inception of the YOLO (You Only Look Once) model in 2015, it has garnered attention for its high recognition accuracy and speed. Currently, the primarily used YOLO models include YOLOv3, YOLOv5, and more recently, YOLOv9 and YOLOv10, which were released in early 2024 and May of this year. This study utilizes YOLOv5, YOLOv9, and YOLOv10 models to detect aircraft in a SAR dataset, comparing their detection performance. The YOLOv10 model demonstrates superiority in both detection accuracy and detection rate. In my research, the YOLOv9 and YOLOv10 shows much more accuracy and frequency in detection.

Keywords: SAR image · YOLOv5 · YOLOv9 · YOLOv10 · Aircraft Target Detection

1 Introduction

In recent years, with the gradual development of SAR (Synthetic Aperture Radar) technology, SAR target detection technology, as the core technology of SAR image processing, has demonstrated significant importance. Aircraft is a typical target in SAR image detection, which is numerous, diverse, and has high observation value. Aircraft detection and recognition based on SAR images can obtain information such as the model, type, and location of aircraft targets, which can effectively assist applications such as dynamic monitoring and situation analysis in key areas. However, traditional target detection methods have struggled to adapt to the numerous issues present in SAR images. In contrast, detection models based on deep learning methods have exhibited strong superiority in image recognition, especially the YOLO series of algorithms. YOLO series of

H.-H. Chen and W. Meng (Eds.): WiSATS 2024, LNICST 606, pp. 64–73, 2025.
https://doi.org/10.1007/978-3-031-86203-8_6

algorithms show high real-time and accuracy in detecting optical image, but it did not perform well in SAR image detection especially aircraft target. However, recently the YOLOv9 and YOLOv10 appear, this research is aimed to test the capability of the newly released model.

2 YOLO Series Detection Algorithm

Although the YOLO series of algorithms exhibit strong advantages in image object detection, earlier versions of YOLO still had numerous shortcomings in SAR image object detection, particularly in complex backgrounds where their feature extraction and generalization abilities were weaker. However, with the release of YOLOv9 and YOLOv10 earlier this year and in late May, they have seen significant improvements in detection accuracy and speed. These updated versions also support the deployment of models on various types of computing power platforms.

2.1 YOLOv5

YOLOv5 is an open-source object detection model that has extensive applications in the engineering field. It utilizes CSPDarknet53 as the backbone network to extract image features, followed by the PAN to fuse the features. Subsequently, it employs preset anchors to initially predict the bounding boxes for object detection. The method of non-maximum suppression (NMS) is adopted to select the most suitable detection boxes. Finally, the network combines the outputs from different prediction heads to obtain the final object detection results.

The network structure of YOLOv5 mainly consists of a backbone network, a neck part[1], and a prediction component. Figure 1 shows its network structure diagram, where Concat is used for feature fusion, Conv1*1 represents a convolution kernel with a size of 1*1, and Conv3*3 represents a convolution kernel with a size of 3*3. In this network structure, the backbone network is primarily used for feature extraction, with the BottleNeckCsp structure integrating the BottleNeck structure and the CSP (Cross Stage Partial) structure, effectively implementing gradient combination. This achieves a reduction in computational complexity and memory requirements while enhancing the learning capabilities of the convolutional neural network. The SPP structure in the network is mainly used to address the issue of inconsistent image scales. The neck part of the network utilizes the Feature Pyramid Network (FPN) and Pyramid Attention Network (PAN) structures. Among them, FPN fuses high-level feature information through up-sampling, while PAN utilizes down-sampling for feature fusion. YOLOv5 combines the use of FPN and PAN structures to better utilize the features obtained from the backbone. The prediction component outputs prediction boxes with categories and confidence scores, mainly including the loss function part and the non-maximum suppression part.

2.2 YOLOv9

The most significant revolution YOLOv9 makes compared to previous generations of networks lies in addressing the issue of data loss during the transmission of data through

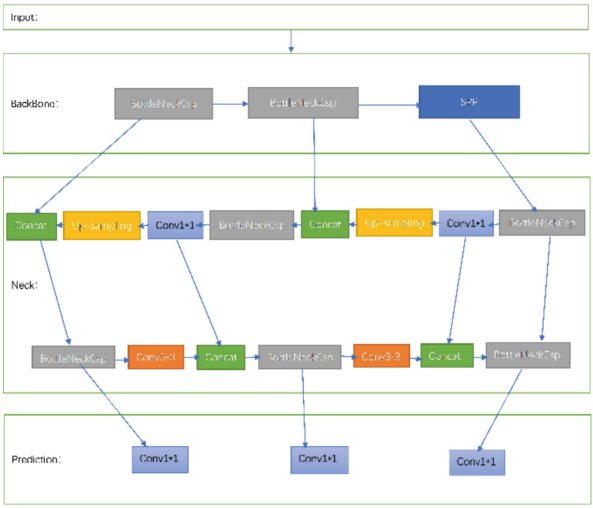

Fig. 1. The Structure of YOLOv5 net

deep learning networks with excessive layers, specifically the loss of significant information when input data undergoes layer-by-layer feature extraction and spatial transformation. To tackle this, YOLOv9 introduces the concept of Programmable Gradient Information (PGI), enabling it to better adapt to the various changes required by deep networks to achieve multiple objectives, thereby obtaining more reasonable values for updating network weights. Another innovation is the introduction of a new network architecture called Generalized Efficient Layer Aggregation Network (GELAN) based on gradient path planning, which achieves better parameter utilization and computational efficiency.

Programmable Gradient Information

In the field of deep networks, the phenomenon of input data losing information during the feedforward process is often referred to as an information bottleneck. Currently, common methods to address this issue include: (1) Employing reversible network architectures. This method primarily leverages repeatedly input data and utilizes deterministic approaches to maintain the information of input data. (2) Utilizing masked modeling. This method mainly relies on reconstruction loss and adopts an implicit approach to maximize extracted features while retaining input information. (3) Introducing deep supervision. This method pre-establishes a mapping between shallow features that have not lost significant information and targets, ensuring that important information can propagate to deeper layers.

However, these methods come with certain drawbacks. For example, reversible architectures require higher inference costs, and to ensure reversibility, the path from the input layer to the output layer cannot be too deep, limiting the training depth. Masked modeling methods can sometimes have incorrect associations between the reconstruction loss and the data. The deep supervision mechanism, due to the stacking of layers, can lead to error accumulation.

To address these problems, YOLOv9 employs the PGI (Programmable Gradient Information) approach. This method generates reliable gradients through auxiliary reversible branches. Since PGI's reversible architecture is built on auxiliary branches, it does not add significant computational costs and is compatible with deeper network structures, making it highly versatile.

The structural diagram of YOLOv9 is shown in Fig. 2[2], which represents the training model of the network. According to the design concept of the PGI module, after training is completed, the auxiliary branches need to be removed for evaluation. Specifically, layers 23 to 27 in the structural diagram, as well as the cv2, cv3, and dfl branches of the detection heads, should be removed to obtain the final model for inference.

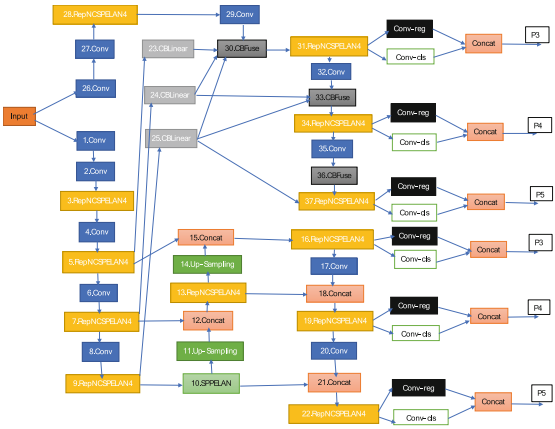

Fig. 2. The Structure of YOLOv9 net

Generalized Efficient Layer Aggregation Network

GELAN (Generalized Efficient Layer Aggregation Network) in YOLOv9 is a lightweight network architecture specifically designed for object detection tasks. GELAN's design takes into account the requirements of real-time applications, aiming to achieve high accuracy while maintaining the network's lightweight nature and high speed.

As the main branch of YOLOv9, GELAN is responsible for feature extraction from the input images. It generates multi-scale feature maps, which incorporate the CSPELAN module and the DOWN module. GELAN introduces the CSPNet structure based on ELAN, achieving the effect of improving computational efficiency. The DOWN module consists of a 2*2 pooling layer and a 3*3 convolution layer, which realizes the down-sampling of feature maps, reducing spatial resolution while increasing the number of channels, facilitating the extraction of multi-scale features. In the design of GELAN, the developers have optimized the network structure through gradient paths, reducing the amount of computation and parameters while maintaining network performance and high accuracy.

The design of GELAN enables YOLOv9 to achieve high accuracy while maintaining lightness and speed, and also provides support for the PGI module.

2.3 YOLOv10

The key improvement made by YOLOv10 lies in its increased detection speed and reduced difficulty in model deployment. The reliance on non-maximum suppression (NMS) in YOLO's post-processing hinders its end-to-end deployment and prolongs the inference time. Additionally, the lack of thorough examination of various component designs in YOLO results in significant computational redundancy, leading to lower computational efficiency. The most significant change in this update is the introduction of a dual label assignment mechanism trained without NMS, which significantly reduces the number of parameters and detection time without compromising recognition accuracy. The structure of YOLOv10 is shown in Fig. 3.

Fig. 3. The structure of YOLOv10 net

Consistent Dual Assignments for NMS Free Training

Unlike one-to-many assignment, one-to-one assignment only assigns one prediction for each label, thereby eliminating the need for non-maximum suppression (NMS) processing[3]. However, this approach can lead to lower supervision intensity, potentially reducing detection accuracy and convergence speed. Nevertheless, this issue can be addressed by utilizing one-to-many matching. To solve this problem, YOLOv10 introduces a dual-label assignment mechanism that combines the advantages of both strategies. Specifically, as shown in Fig. 4, YOLOv10 adds a one-to-one prediction head. The one-to-one branch retains the same structure as the one-to-many branch and adopts the same optimization objectives, but utilizes one-to-one assignment for label assignment. This means that the one-to-one and one-to-many branches are structurally consistent, sharing the same network architecture and optimization strategies, but employing different matching strategies when determining label assignments. The one-to-one branch

ensures that each prediction is associated with only one ground truth through one-to-one assignment, while the one-to-many branch may associate multiple predictions with the same ground truth, providing a stronger supervision signal. During training, the two prediction heads and the model are jointly optimized, allowing the backbone and neck networks to leverage the benefits of both allocation strategies. At inference time, the one-to-many prediction head is discarded, and only the one-to-one predictions are utilized, reducing the complexity of network deployment.

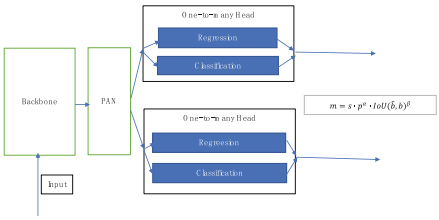

Fig. 4. Consistent dual assignments for NMS-free training

3 Experimental Result and Analysis

3.1 Description of Dataset

The experimental dataset is SAR-AIRcraft-1.0, which comes from the GF-3 satellite, adopts single polarization, has a spatial resolution of 1m, and the imaging mode is spotlight mode. Considering the size of the airport and the number of aircraft, the dataset mainly selects image data from Shanghai Hongqiao Airport, Beijing Capital Airport, and Taiwan Taoyuan Airport, including four sizes of 800*800, 1000*1000, 1200*1200, and 1500*1500. It contains 4368 photos and 16463 aircraft instances. The aircraft categories include seven categories: A220, A320/321, A330, ARJ21, Boeing 737, Boeing 787, and other.[4] The number of each category is shown in the Fig. 5.The dataset has the following characteristics:

1) Complex scenes: The dataset includes images from multiple civil airports at different times, covering a large area. The background includes facilities such as terminal buildings, vehicles, and buildings, which are highly complex.
2) Rich categories: The dataset contains fine-grained category information for different aircraft targets, and the similarity of scattering characteristics between different categories increases the difficulty of recognition.
3) Dense targets: The targets are densely distributed in the airport.
4) Noise interference: Due to SAR imaging characteristics, there is speckle noise interference in the image.
5) Multiple scales: Some aircraft targets are less than 50*50 in size, while some targets are larger than 100*100.

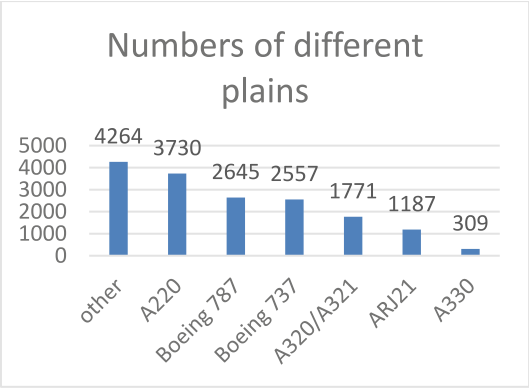

Fig. 5. Number of different plain

3.2 Training and Analysis

The experimental model in this document is based on the Windows 11 platform, constructed under the PyTorch framework, and operated within Anaconda. The hardware environment is shown in Table 1.

Table 1. Hardware environment

Heading level	Example
CPU	14700 K
RAM	64 GB
GPU	Nvidia RTX4070tisuper
GPU Mem	16 GB

In the translation experiment, models such as YOLOv5, YOLOv9, YOLOv9-e, YOLOv10-n, and YOLOv10-l were selected. The main parameters for training the models are presented in Table 2.

The time consumed for training the aforementioned models is shown in Fig. 6. Yolov5 takes the shortest time, with an average of 24 s per batch for training. Followed by Yolov10n, which requires 58 s per batch for training. Yolov10l needs 86 s per batch, while Yolov9 takes 118 s per batch. Yolov9-e has the longest training time, requiring 1573 s per batch.

The number of FLOPs of the above models is shown in Fig. 7, where Yolov10n requires the least number of floating-point operations, needing only 8.4 GFLOPs. Followed by Yolov5m, which requires 48.3 GFLOPs. Yolov10l needs 127.2 GFLOPs, Yolov9-e requires 244.9 GFLOPs, and Yolov9 has the highest demand, needing 266.2 GFLOPs (Fig. 8).

Table 2. parameters of selected models.

Model	Epoch	Image size	Batch size	Learning rate
yolov5	100	640	16	0.01
yolov9	100	640	16	0.01
yolov9-e	100	640	16	0.01
yolov10-n	100	640	16	0.01
Yolov10-l	100	640	16	0.01

Fig. 6. Training time

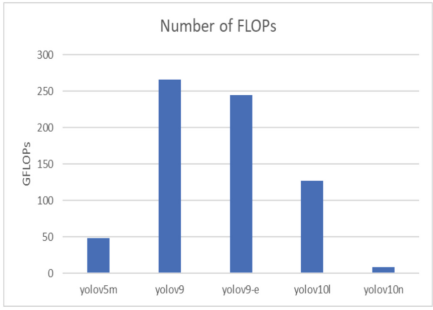

Fig. 7. Number of FLOPs

The amount of GPU memory required indicates the difficulty of deploying a model on various platforms. Models that require more GPU memory tend to have a higher deployment difficulty. The GPU memory usage of each model is shown in Fig. 9. Among them, Yolov10n has the lowest GPU memory usage, requiring only 3.2 GB. Followed by Yolov5 m, which needs 7.38 GB. Yolov10l requires 14.2 GB, Yolov9 needs 17.2 GB, and Yolov9-e demands the most, requiring 24.1 GB.

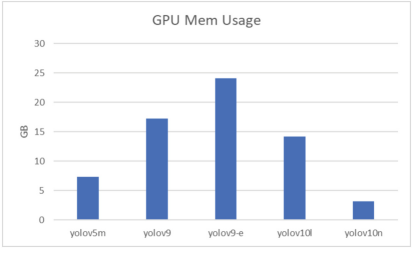

Fig. 8. GPU Memory Usage

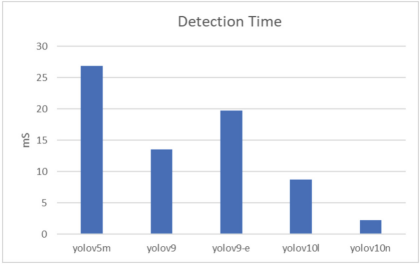

Fig. 9. Detection time

The detection time indicates the time required by the model to process actual images. The shorter the detection time is, the higher the real-time performance of the model for recognition. The detection time is shown in Fig. 9.

In testing the recognition accuracy of a trained model, one of the most crucial metrics is the mAP50–95 parameter. We have collected the mAP50–95 values of the trained models for detecting target objects at different epochs. This data reflects the changes in the model's recognition accuracy as the training process progresses. A model with faster increases in mAP50–95 and higher values as the number of training epochs increases can be considered more precise and having stronger learning capabilities. Using the collected information, we have plotted the curves for different models, as shown in Fig. 10.

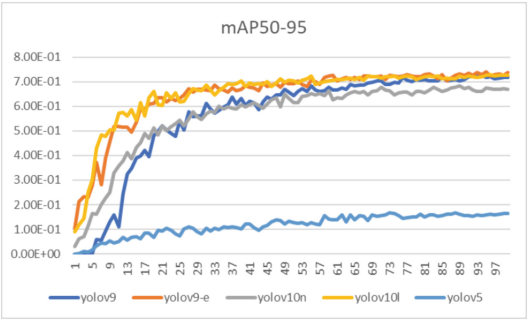

Fig. 10. mAP50–95 of different model.

We can observe that the recognition accuracy and learning ability of yolov9, yolov9-e, yolov10l, and yolov10-n are quite similar, and all four of them significantly outperform the yolov5 model, making them more suitable for aircraft target detection tasks in SAR images. Figure 11 to Fig. 13 demonstrates the effects of aircraft target detection in the SAR dataset using models trained with yolov10l and yolov5, where Fig. 11 is the labeled picture. Figure 12 is the detection result of yolov5m, Fig. 13 is the detection result of yolov10l. From the figure, we can observe that the model trained with yolov10l is able to accurately identify almost every aircraft target on the tested SAR images, while yolov5 is nearly incapable of performing such recognition.

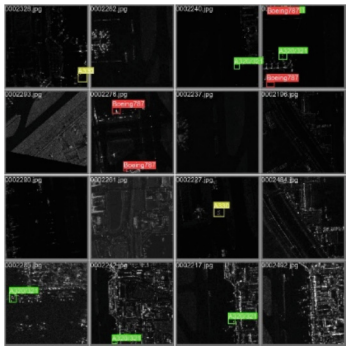

Fig. 11. Labelled picture

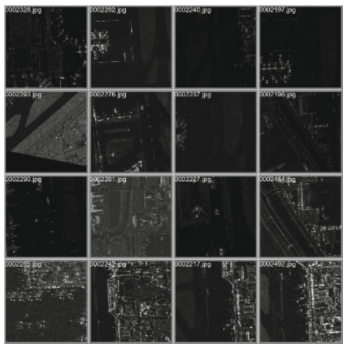

Fig. 12. Predition by yolov5m

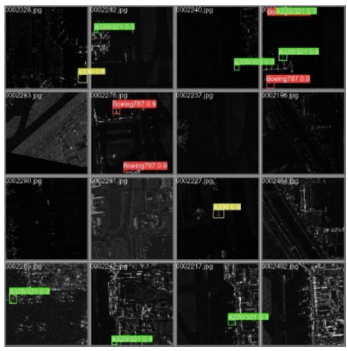

Fig. 13. Predition by yolov10l

4 Conclusion

Based on the result and analysis presented in the study above, the conclusions obtained are as follows:

1) Models trained by yolov9, yolov9-e, yolov10l, and yolov10n all achieve high detection accuracy for aircraft targets in SAR images. Among them, yolov9-e achieves the highest accuracy but requires significantly longer training and recognition time compared to the other three models. The yolov10n model, while having relatively lower accuracy, boasts the fastest training and recognition speed with lower deployment difficulty, making it suitable for deployment on small platforms. Additionally, yolov9 and yolov10l exhibit similar performance, both demonstrating excellent performance in aircraft target recognition tasks.
2) Due to the poor imaging quality of SAR images, lack of geometric correction, and the similarity in SAR imaging results of aircraft targets, it is difficult for human eyes to accurately identify and distinguish different types of aircraft targets on SAR images. However, the new generation of YOLO models has demonstrated strong recognition performance, achieving high accuracy while also maintaining speed, and accurately distinguishing aircraft categories.
3) The current new generation of Yolo object detection models have strong practical application potential in the field of aircraft target detection in SAR, capable of meeting the requirements of high accuracy, multi-platform, and high speed.

References

1. Tian, M., Liao, Z.: 2021, Research on flower image classification method based on YOLOv5. J. Phys. Conf. Ser. **1**, 012022 (2024). https://doi.org/10.1088/1742-6596/2024/1/012022(2024)
2. Wang, C.Y., Yeh, I.H., Mark Liao, H.Y.: YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv preprint* arXiv:2402.13616 (2024)
3. Wang, A., et al.: YOLOv10: Real-Time End-to-End Object Detection. *arXiv preprint* arXiv: 2405.14458 (2024)
4. Zhirui, W., Yuzhuo, K., Xuan, Z., et al.: SAR-AIRcraft-1.0: High-resolution SAR aircraft detection and recognition dataset. J. Radars **12**(4), 906–922 (2023). https://doi.org/10.12000/JR23043