EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems

EAI.EU Research Article

Efficiency Cost-Sensitive Loss of Transformer based on Mamba Mechanism for Aircraft Detection in Satellite Imagery

Manh-Tuan Do[1], Manh-Hung Ha[2,*], Minh-Huy Le[3], Oscal Tzyh-Chiang[2,4]

[1]Faculty of Sciences and Technology, UEVE, University of Paris Saclay, Évry-Courcouronnes 91000, France
[2]Faculty of Applied Sciences, International School, Vietnam National University, Hanoi 100000, Vietnam
[3]Faculty of Electrical and Electronic Engineering, Phenikaa University, Yen Nghia, Hanoi, 100000, Hanoi, Vietnam
[4]Department of Electrical Engineering, National Chung Cheng University, Chiayi, Chiayi, 62102, Chiayi, Taiwan

Abstract

Detecting aircraft in satellite images poses considerable challenges due to complex backgrounds and variable conditions influenced by sensor geometry and atmospheric factors. Despite rapid advancements in deep learning algorithms, their main focus has been on ground-based imagery. This study offers a thorough evaluation and comparison of advanced object detection algorithms specifically designed for aircraft detection in satellite imagery. By leveraging the extensive HRPlanesV2 dataset and a rigorous validation process on the GDIT dataset, we trained a cutting-edge object detection model, YOLO-Mamba, published in June 2024. Additionally, we introduce YOLO-Mamba-TransGhost, which integrates a novel Transformer module SC3T and Ghost Convolution into the YOLO model's backbone architecture. Furthermore, substituting the WIoU loss function with CIoU in YOLO-Mamba results in significant improvements in accuracy and small object detection. Experimental results on the GDIT dataset indicate that YOLO-Mamba-TransGhost improves mAP@.5 by approximately 2% compared to the original YOLO-Mamba. Similarly, tests on the HRPlanev2 data set reveal a notable reduction in model complexity and an impressive accuracy of 98.7% which is achieved by leveraging a cost-sensitive loss function that dynamically focuses training on higher quality samples, improving convergence and accuracy.Therefore, the proposed YOLO-Mamba-TransGhost model demonstrates superior accuracy and reduced complexity in aircraft detection from satellite imagery, highlighting its potential for practical applications in aerospace monitoring, disaster management, and surveillance systems domain.

Received on 23 December 2024; accepted on 14 September 2025; published on 19 September 2025

Keywords: YOLO, Mamba, Transformer, Ghost Convolution, Wiou loss function

doi:10.4108/eetinis.v12i4.8279

1. Introduction

In recent years, advancements in remote sensing technology have significantly enhanced the quality and richness of satellite imagery, making these images an indispensable asset across a wide range of applications, especially in military operations. This has led to object detection in remote sensing becoming a critical research area, with particular focus on aircraft detection due to its crucial role in airport monitoring, military reconnaissance, and strategic decision-making. However, detecting aircraft in satellite imagery remains challenging due to factors such as the small size of targets, the high-altitude acquisition of images, and various environmental influences, including weather conditions, illumination, and sensor-specific parameters. Moreover, the dense arrangement of aircraft within scenes makes it difficult to separate them from the background, complicating feature extraction and reducing detection accuracy, which hinders real-time detection performance.

[*]Corresponding author. Email: hunghm@vnu.edu.vn

The advent of deep learning has revolutionized object detection, offering major improvements over traditional machine learning approaches. Current object developers fall into two main categories: two-stage models, such as R-CNN [1], Fast R-CNN [2], and Faster R-CNN [3], which prioritize accuracy; and one-stage models, like SSD [4] and the YOLO family [5], which emphasize speed and real-time performance.

Despite considerable progress in detection accuracy and speed for satellite imagery, significant challenges remain. Initially, YOLO v1 [6] had a large positioning error and a lower recall rate compared to region-based proposal methods like Fast R-CNN. After that, various studies have since proposed improvements to these models. The main enhancements of YOLOv2 [7] include improving the recall rate, batch normalization, anchor boxes, and multi-scale training.

Further developments, such as incorporating DenseNet into YOLOv3, improved detection accuracy at the expense of increased complexity and reduced speed [8]. Other approaches include lightweight modifications to YOLOv3 to balance accuracy and speed [9], and the integration of advanced modules and activation functions in YOLOv4, YOLOv5, and YOLOv6 for performance improvements [10]-[12]. More recently, YOLOv7 incorporated the channel attention mechanism from CBAM, refining its network architecture. Additionally, it replaced complete intersection over union (CIoU) with Alpha-GIoU as the coordinate loss function, leading to improved generalization capabilities.

In 2023, YOLOv8 was enhanced with Transformer blocks to boost accuracy, particularly for detecting small objects [13]. Early in 2024, YOLOv9 was developed and applied to detect aircraft in images from Low Earth Orbit (LEO) satellites [14], demonstrating its effectiveness in real-world scenarios.

In May 2024, YOLOv10 [15] was launched, featuring innovations such as Consistent Dual Assignment and Holistic Efficiency-Accuracy Driven Model Design, which reduces the reliance on Non-Maximum Suppression, thereby minimizing redundant information and decreasing latency. These advancements also contributed to a significant improvement in accuracy.

Nevertheless, challenges remain when detecting aircraft in remote sensing imagery due to factors such as low resolution, complex backgrounds, varying object sizes and orientations, and imbalanced datasets.

In this study, we evaluate and compare several cutting-edge object detection algorithms designed specifically for aircraft detection in satellite imagery. Moreover, we assess the performance of YOLO-Mamba [16], a state-of-the-art object detection model introduced in June 2024. We also propose an enhanced version, YOLO-Mamba-TransGhost, which integrates the SC3T Transformer and GhostConV into its backbone to improve accuracy. Additionally, we recommend replacing the WIoU loss function [17] with CIoU in YOLO-Mamba, which offers an intelligent gradient gain allocation strategy. This strategy reduces the competitiveness of high-quality anchor boxes while mitigating the adverse gradient impact from low-quality samples. By focusing on mid-quality anchor boxes, it improves the overall performance of the detector.

The structure of this paper is as follows: Section II introduces the YOLO-Mamba-TransGhost model and details its design process; Section III outlines the experimental data, environment, and comparative results; and Section IV presents the conclusions and future directions. With these advancements, we aim to address the inherent challenges of remote sensing image analysis, contributing to the broader field of computer vision.

2. METHOD

2.1. Proposed YOLO-Mamba-TGW

In general, YOLO-Mamba, developed as an advanced version of the YOLO series, represents a significant leap in real-time object detection. This model integrates several innovations aimed at overcoming the limitations of earlier YOLO versions. The backbone architecture incorporates State Space Models (SSMs), which significantly reduce the computational complexity associated with attention mechanisms while preserving the ability to model long-range dependencies. The ODSSBlock, a core innovation, utilizes a multi-directional scan mechanism (SS2D) to enhance spatial information capture, allowing for more accurate object detection in complex environments. Additionally, the model includes the LSBlock, which leverages depthwise separable convolutions to balance computational efficiency with high processing performance. The RGBlock uses a gating mechanism to optimize information flow, ensuring the effective capture of both global and local features, even in challenging detection scenarios. These innovations allow YOLO-Mamba to perform exceptionally well on large-scale datasets such as COCO and VOC, achieving higher accuracy (mAP) and better computational efficiency (FLOPs) compared to previous YOLO versions, while maintaining real-time detection capabilities.

We propose an advanced variant, YOLO-Mamba-TGW where TGW refers to TransGhost with WIoU, the proposed YOLO-Mamba-TransGhost model integrating the SC3T Transformer, Ghost Convolution, and WIoU loss function, as illustrated in Figure 1. The diagram illustrates the architecture of the proposed, starting from the input image through a multi-stage process. The model includes an ODMamba Backbone for feature extraction, followed by the PAFPN Neck which merges multi-scale features using upsampling and concatenation. Notably, modules like GhostConv, ODSSBlock,

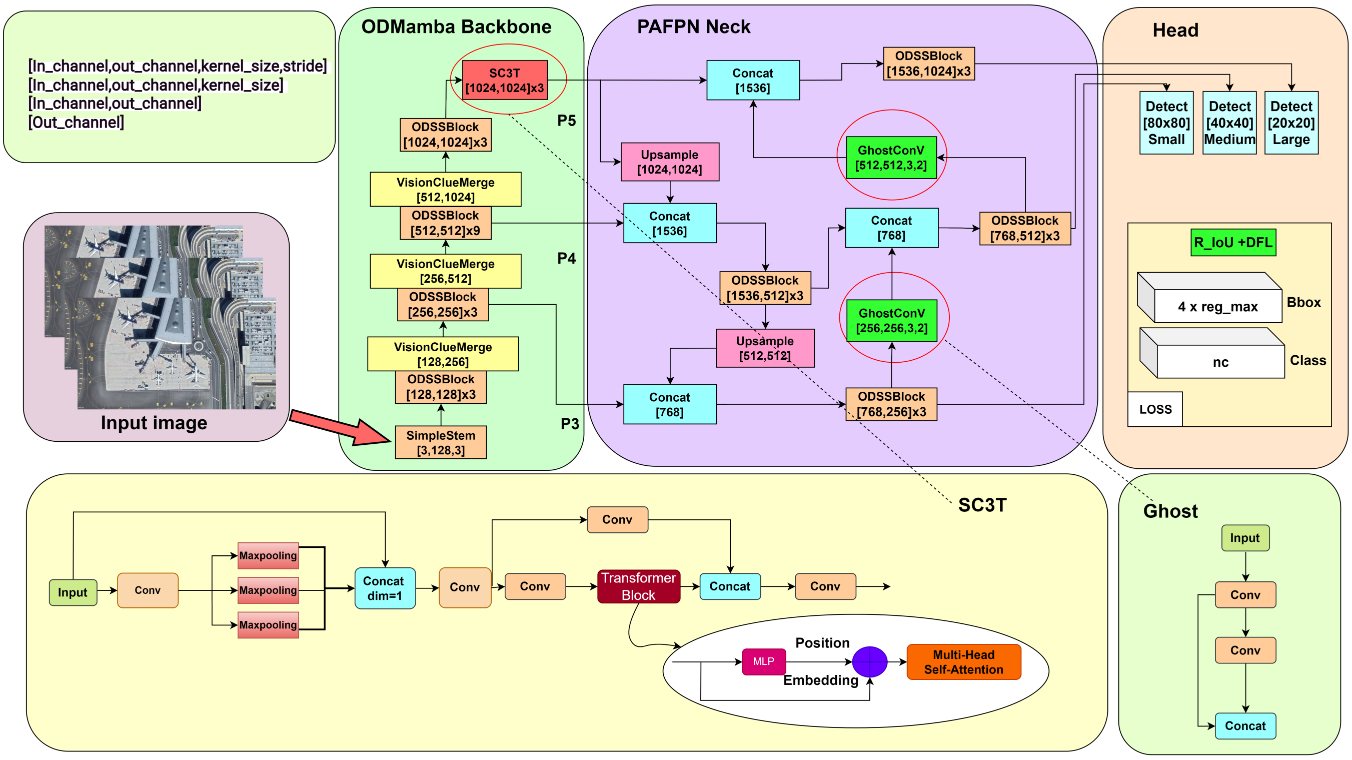


Figure 1. The architecture of YOLO-Mamba-TGW.

and SC3T (with transformer blocks and multi-head self-attention) are integrated to enhance performance and reduce computation. The final Head performs detection at three scales (small, medium, large) with bounding-box regression and classification using R_IoU + DFLloss. This architecture balances accuracy and efficiency, suitable for complex aerial image analysis. This model extends the backbone architecture by incorporating the SC3T module, a Transformer-based component that enhances the capture of global features and increases the receptive field. This extension improves the model's capacity to process large-scale and complex datasets by aggregating high-level contextual information. Furthermore, we replace the traditional convolution layers in the head section with Ghost convolution, significantly reducing the computational cost. Ghost convolution generates a subset of feature maps and expands these through linear operations, thereby reducing both the number of parameters and computational complexity, while maintaining the detection accuracy.

In addition, we substitute the WIoU loss function with CIoU in YOLO-Mamba, which results in notable improvements in accuracy, particularly in detecting small objects. CIoU provides better optimization by considering overlap area, distance, and aspect ratio between predicted and ground truth boxes, thus improving overall object localization.

The SC3T module [18] combines a Spatial Pyramid Pooling (SPP) structure with a C3TR module to handle varying input image sizes and multiscale feature extraction. The SPP module utilizes kernels of different sizes (5×5, 9×9, 13×13) to aggregate multi-scale features, and the resulting feature maps are concatenated across channels, enhancing the model's ability to process features from different scales. The C3TR module incorporates a Transformer block at the three outputs of the detection network, which employs a Multi-Head Self-Attention (MSA) mechanism. The MSA captures global contextual information by updating and concatenating Query (Q), Key (K), and Value (V) representations from different spatial regions, ensuring the model effectively captures global features. These representations are then passed through a Multilayer Perceptron (MLP) to refine the output and improve feature expression.

In order to further optimize the efficiency of the proposed YOLO-Mamba-TGW, we integrate *Ghost convolution* into the head of the network. Traditional convolutional layers tend to produce redundant feature maps, as many of them contain overlapping information across different channels. This redundancy leads to unnecessary computational overhead. Ghost convolution [19] addresses this issue by generating only a subset of the feature maps through standard convolution operations. These initial feature maps are then expanded using linear transformations, which produce additional feature maps without the need for full convolutions, thereby significantly reducing the computational load and the number of parameters required.

Mathematically, the process of generating feature maps in a convolutional layer can be described as:

$$Y = X \cdot f + b \tag{1}$$

where $X \in \mathbb{R}^{c \times h \times w}$ represents the input feature map, with c as the number of input channels, and h and w as the height and width, respectively. The output feature map $Y \in \mathbb{R}^{h' \times w' \times n}$ has n channels, while $f \in \mathbb{R}^{c \times k \times k \times n}$ denotes the convolutional filters with a kernel size $k \times k$, and b is the bias term. The computational complexity for this operation is given by:

$$C_{\text{std}} = n \cdot h' \cdot w' \cdot c \cdot k \cdot k \tag{2}$$

Ghost convolution improves efficiency by introducing a two-step process. First, a subset of feature maps is created using the standard convolution operation. Then, additional feature maps are generated through lightweight linear operations. The total computational cost can be expressed as

$$C_{\text{ghost}} = \frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot r \cdot r \tag{3}$$

Here, s represents the redundancy ratio specified as s is greater than one (typically 2 in our experiments), and $r \times r$ is the kernel size for the linear operations. This technique significantly reduces the overall computational cost when compared to traditional convolutions, as the Ghost module requires only a fraction of the operations needed in full convolutions. In fact, the total computational cost using Ghost convolution is approximately $1/s$ of the cost of a regular convolution, making it a highly efficient method for reducing complexity without sacrificing the quality of the features.

By applying this method, YOLO-Mamba-TGW achieves a considerable reduction in the number of operations and parameters required for convolutional layers, which directly translates to improved processing speed and efficiency. This is particularly beneficial for real-time object detection tasks, where both accuracy and computational efficiency are critical.

2.2. Improved Loss Function

The YOLO-Mamba model initially employed the Complete IoU (CIoU) loss function, which was designed to enhance detection accuracy by considering three factors: the distance between the centers of the predicted and ground-truth bounding boxes, the difference in their aspect ratios, and the ratio of their diagonal distances. This approach was particularly effective for improving the detection of smaller objects by providing better localization and bounding box adjustments.

In this study, we propose replacing CIoU with an improved version of the Weighted IoU (WIoUv3) loss function for bounding box regression that embodies a cost-sensitive approach, a key concept for understanding the title of this paper. Unlike standard loss functions that treat all detection errors equally, a cost-sensitive method applies different 'costs' or weights to different training examples based on their quality. WIoUv3 achieves this through a dynamic, non-monotonic focusing mechanism that adjusts the gradient gain for each anchor box. By effectively reducing low quality samples that can produce harmful gradients and focusing training on 'ordinary quality' ex amples that are most beneficial for learning, 'sensitivity' to sample quality leads to more stable convergence and better detection accuracy. This new version builds on the previous WIoUv1 and WIoUv2, both of which focused on refining the model's ability to balance simple and difficult examples. WIoUv1 incorporated attention mechanisms that prioritized distance metrics, while WIoUv2 introduced a monotonic focus coefficient to further optimize gradient distribution. WIoUv3 enhances this methodology by dynamically scaling the gradient gain according to the anchor box quality, allowing for more precise adjustments during training.

The WIoUv3 loss function is defined as:

$$L_{WIoUv3} = r \times L_{WIoUv1} \tag{4}$$

where r is the non-monotonic focus coefficient, calculated as:

$$r = \frac{\beta}{\delta \alpha \beta^{-\delta}} \tag{5}$$

In this formulation, β represents the outlier factor, computed as:

$$\beta = \frac{L^*_{IoU}}{L_{IoU}}, \quad \beta \in [0, +\infty) \tag{6}$$

Here, L_{WIoUv1} represents the original WIoUv1 loss, which is calculated as:

$$L_{WIoUv1} = R_{WIoU} \times L_{IoU} \tag{7}$$

with R_{WIoU} being the weighting factor:

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{w^2 + h^2}\right) \tag{8}$$

and L_{IoU}, the standard IoU-based loss, being defined as:

$$L_{IoU} = 1 - IoU \tag{9}$$

The primary strength of WIoUv3 lies in its ability to dynamically focus on anchor boxes with ordinary or lower quality, scaling down the gradient gain for higher-quality boxes. This prevents the model from being influenced by potentially harmful gradients that originate from low-quality samples. As a result, the loss function promotes better convergence and overall detection performance, particularly in scenarios involving small objects or imbalanced datasets.

We chose to integrate WIoUv3 into the YOLO-Mamba model due to its capacity to adapt to varying qualities of sample data. This flexibility ensures more accurate localization, faster convergence, and improved robustness, especially in challenging detection tasks. By dynamically adjusting its focus, WIoUv3 plays a critical role in helping the model generalize effectively across a wide range of object detection scenarios, making it an essential component of our proposed improvements.

3. EXPERIMENTAL ENVIRONMENT AND DATASETS

3.1. Experiment Environment

The proposed YOLO-Mamba-TGW model was trained on the GDIT dataset and HRPlanev2 dataset using Google Colab with a High RAM environment and a Tesla V100 GPU. After the training process was completed, the corresponding weight sets for each model were generated. The models were then evaluated on their respective test datasets. Finally, the results were compared to assess the improvements brought by YOLO-Mamba-TGW.

3.2. Datasets

GDIT Dataset

The GDIT Aerial Airport dataset [20] is a curated collection of aerial images focused on parked aircraft within airport settings. All types of aircraft in this dataset are categorized under a single label, 'airplane.' This dataset serves as a valuable resource for researchers developing and evaluating aircraft detection algorithms. It consists of 338 high-resolution images, each measuring 600 × 600 pixels, which are split into training, validation, and testing sets containing 236, 68, and 34 images, respectively. The dataset is further enhanced with augmented training images that include variations in filters, zoom levels, and rotations, expanding the total number of images to 810. By providing a single classification label for all aircraft types, the GDIT dataset streamlines the detection process, aiding in the creation of effective detection models. The GDIT dataset was chosen as a curated benchmark for evaluating the core accuracy of aircraft detection models in a controlled airport environment. Its high-resolution images and single-class focus allow for a precise assessment of localization performance.

HRPlanesv2 Dataset

HRPlanev2 dataset [21], which is designed for high-resolution airplane detection tasks. The HRPlanev2 dataset comprises imagery sourced from prominent global airports and aircraft boneyards, offering a diverse range of conditions in terms of landscape, seasonal variations, and data acquisition geometry. The images were obtained from Google Earth and cover significant locations including Paris-Charles de Gaulle, John F. Kennedy, Frankfurt, Istanbul, Madrid, Dallas, Las Vegas, Amsterdam Airports, and Davis-Monthan Air Force Base.

The dataset includes 3092 RGB images, each with a resolution of 4800 x 2703 pixels. These images have been manually annotated with bounding boxes around each aircraft, utilizing HyperLabel software for the annotation process. A rigorous quality control procedure was implemented, involving visual inspections by independent analysts not involved in the initial annotation, ensuring the accuracy of the labels. In total, the dataset includes 18,477 annotated aircraft. which are very large images, diverse geographical locations, varied seasonal conditions, and a sheer number of annotations provide a comprehensive test of the model's generalization capacities.

The HRPlanev2 dataset is divided into three subsets for experimental purposes: 70% (2166 images) for training, 20% (615 images) for validation, and 10% (311 images) for testing. This distribution facilitates robust model training and evaluation, enabling a thorough assessment of detection performance across a range of conditions and scenarios.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Model Performance Evaluation

In this section, we thoroughly evaluate the performance of several state-of-the-art object detection models, including RT-DETR, YOLOv3s-tiny, YOLOv5s, YOLOv6, YOLOv8-World, YOLOv10s, YOLO-Mamba, and the advanced YOLO-Mamba-TransGhost. The models were benchmarked on two distinct datasets, GDIT and HRPlanev2, using key performance metrics such as Precision, Recall, mean Average Precision at IoU thresholds of 0.5 and 0.95 (mAP@.5, mAP@.95), model complexity (in terms of Parameters), and GFLOPS. The experimental results are summarized in Table 1.

On the GDIT dataset, YOLO-Mamba-TransGhost performed best in terms of Precision (92.20%), Recall (87.70%), and mAP@.5 (94.30%). While RT-DETR had a slightly higher Recall (89.90%), it came with a lower Precision (91.40%), indicating more false positives. YOLO-Mamba-TransGhost also led in mAP@.95 with 51.20%, demonstrating superior localization accuracy. Although YOLO-Mamba and YOLOv10s followed closely, the improvements brought by the TransGhost module are evident, particularly in challenging detection tasks.

However, this increase in performance comes with higher computational costs. YOLO-Mamba-TransGhost requires 33.7M parameters and 102.2 GFLOPS, making it more resource-intensive compared to YOLO-Mamba and YOLOv10s. This may present challenges in resource-limited environments.

To further substantiate the effectiveness of the proposed YOLO-Mamba-TransGhost model, we tested it on the HRPlanev2 dataset. Similar to the GDIT dataset, the results on HRPlanev2 reaffirm the model's superior performance. YOLO-Mamba-TransGhost again showed the best performance with 98.70% Precision, 96.00% Recall, and 99.20% mAP@.5. It also achieved the highest mAP@.95 of 80.50%, highlighting its effectiveness in precise detection, especially in aerial imagery. Despite the higher complexity, the improvements in accuracy and detection make YOLO-Mamba-TransGhost an excellent choice for scenarios where accuracy is critical.

In summary, the proposed YOLO-Mamba-TransGhost model consistently demonstrates superior performance on both the GDIT and HRPlanev2 datasets. Its notable improvements in Precision, Recall, and mAP@.5/mAP@.95 metrics highlight the effectiveness of the TransGhost module in enhancing both object classification and localization tasks. Although the model's computational complexity is higher, the significant gains in accuracy, particularly for high-precision applications like satellite imagery analysis, make it a compelling solution for scenarios where accuracy is prioritized over computational efficiency.

4.2. Loss Function Comparison

To further improve the detection accuracy of the YOLO-Mamba-TransGhost model, we experimented with various loss functions, including CIoU, DIoU, EIoU, GIoU, SIoU, and WIoU. The comparative results for these loss functions, including their respective mAP@.5 and number of parameters, are presented in Table 2.

The WIoU loss function demonstrated the best performance with a mAP@.5 score of 94.30%, clearly outperforming the other variants. The second highest result came from CIoU, with a mAP@.5 of 93.90%, followed closely by EIoU, which achieved 93.50%. These results show that while CIoU and EIoU are quite effective, WIoU offers superior object detection accuracy.

Interestingly, the DIoU loss function yielded a comparatively lower mAP@.5 of 92.90%, and GIoU had the lowest score of 91.80%. These scores suggest that DIoU and GIoU, though popular for their efficiency in bounding box regression, may not be as suitable for this particular task compared to the other functions.

It's also worth noting that all these models have identical numbers of parameters (33.7M), ensuring that the variations in detection performance are purely attributable to the different loss functions rather than architectural changes.

By comparing these results with earlier experiments, it is clear that the adoption of the WIoU loss function provides a noticeable advantage in object detection for the YOLO-Mamba-TransGhost model, achieving the highest detection accuracy with 94.30%. This result reinforces the importance of selecting an appropriate loss function to enhance model performance, particularly for challenging tasks such as object detection in complex datasets like GDIT.

4.3. Ablation

Our ablation study provides valuable insights into the enhancements of YOLO-Mamba-TransGhost as shown in Tabel 3. The baseline YOLO-Mamba B, using Convolution (ConV) with the CIoU loss function, achieved an mAP of 92.50% with 21.7M parameters. Introducing Transformer modules improved the mAP to 92.70%, although this increase also raised the parameter count to 34M. Switching to Depthwise Convolution (DWConv) further improved mAP to 93.00% while keeping the parameter count at 33.5M. The most significant accuracy gain was observed with Ghost Convolution (GhostConV), which achieved a high mAP of 93.90% with 33.7M parameters.

Additionally, replacing the CIoU loss function with WIoU resulted in the highest mAP of 94.30%. These results confirm that combining GhostConV with WIoU provides the most substantial improvements in detection accuracy, demonstrating the effectiveness of YOLO-Mamba-TransGhost with WIoU Loss Function in satellite imagery analysis.

4.4. Compared to other studies

Based on the comparison of parameters and GFLOPS in Table 1, models of YOLOv6 and YOLOv8-World show strong potential for deployment on devices with limited hardware because of using less memory and faster inference time (only 4.2M and 4M parameters and low GFLOPS). On the other hand, YOLO-Mamba-TGW has a larger size (33.7M parameters) and higher computational cost (102.2 GFLOPS), but it delivers the best performance, reaching the highest mAP@0.5 (99.2%) and mAP@0.95 (80.5%) on the HRPlanev2 dataset. This means YOLO-Mamba-TGW is a suitable option when high accuracy is needed and hardware resources are available. Therefore, choosing the right model depends on carefully balancing speed, memory efficiency, and accuracy

Table 1. Performance Comparison of Object Detection Models

Datasets	Model	Precision	Recall	mAP@.5	mAP@.95	Parameters	GFLOPS
GDIT	RT-DETR	91.40%	89.90%	93.90%	51.00%	32M	103.4
	YOLOv3s-tiny	89.90%	80.00%	87.40%	46.00%	12.1M	18.9
	YOLOv5s	93.60%	86.20%	91.80%	50.50%	9.1M	22.8
	YOLOv6	91.00%	85.50%	91.80%	49.30%	4.2M	11.8
	YOLOv8-World	91.60%	85.20%	91.90%	49.20%	4M	12.8
	YOLOv10s	89.10%	85.20%	92.20%	49.90%	8M	24.4
	YOLO-Mamba	91.90%	85.70%	92.50%	50.10%	21.8M	49.6
	YOLO-Mamba-TGW	92.20%	87.70%	94.30%	51.20%	33.7M	102.2
HRPlanev2	YOLOv10s	95.90%	93.40%	97.30%	76.40%	8M	24.4
	YOLO-Mamba	97.10%	95.30%	98.10%	77.10%	21.7M	49.7
	YOLO-Mamba-TGW	98.70%	96.00%	99.20%	80.50%	33.7M	102.2

Table 2. Comparison of different loss functions on the YOLO-Mamba-TransGhost model.

Model	Loss Function	mAP@.5	Parameters
YOLO-Mamba-TransGhost	CIoU	93.90%	33.7M
YOLO-Mamba-TransGhost	DIoU	92.90%	33.7M
YOLO-Mamba-TransGhost	EIoU	93.50%	33.7M
YOLO-Mamba-TransGhost	GIoU	91.80%	33.7M
YOLO-Mamba-TransGhost	SIoU	93.20%	33.7M
YOLO-Mamba-TransGhost	**WIoU**	**94.30%**	**33.7M**

Table 3. Performance comparison of YOLO-Mamba B with different settings.

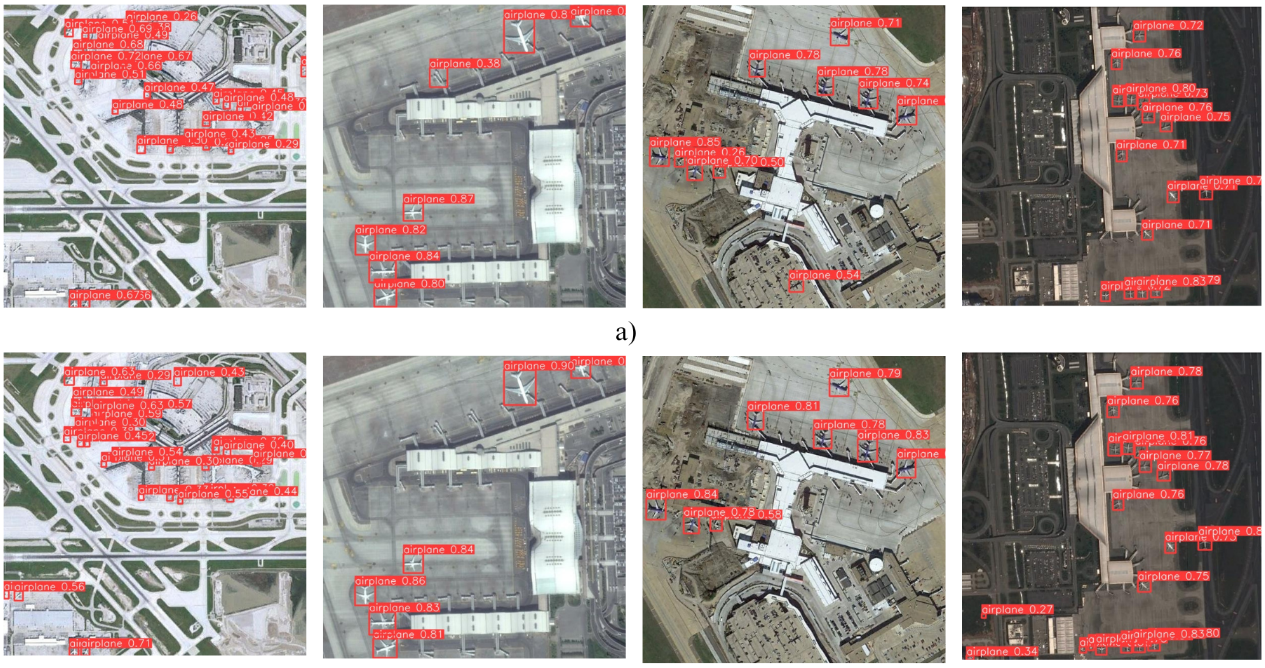
Base line	Transformer	ConV type	Loss Function	mAP@.5	Parameters
YOLO-Mamba B	No	ConV	CIoU	92.50%	21.7M
YOLO-Mamba B	Yes	ConV	CIoU	92.70%	34M
YOLO-Mamba B	Yes	DWConv	CIoU	93.00%	33.5M
YOLO-Mamba B	Yes	GhostConV	CIoU	93.90%	33.7M
YOLO-Mamba B	**Yes**	**GhostConV**	**WIoU**	**94.30%**	**33.7M**

Table 4. Comparison with Other Studies

References	Models	mAP@.5
Safouane EL GHAZOUALI [22]	RetinaNet	81.90%
	SSD	59.40%
	Faster RCNN	57.30%
	RTMDet	86.30%
	YOLOv8	91.90%
Joel Bhaskar Nadar [23]	YOLO-NAS S	81.50%
	YOLO-NAS M	81.80%
Our experiments	YOLO-Mamba	92.50%
	YOLO-Mamba-TGW	94.30%

based on the deployment needs In the context of the GDIT dataset, our experiments with YOLO-Mamba-TransGhost demonstrate a notable advancement over previously established object detection models as shown in Table 4. For comparison, Safouane EL GHAZOUALI's study [22] assessed several leading models for aircraft detection in satellite imagery, including RetinaNet, SSD, Faster RCNN, RTMDet, and YOLOv8. Among these, YOLOv8 achieved an impressive mAP@.5 of 91.90%, whereas SSD and Faster RCNN performed less effectively, with mAP@.5 scores of 59.40% and 57.30%, respectively. The RTMDet model reached an mAP@.5 of 86.30%, highlighting its capability in handling intricate detection tasks.

Our YOLO-Mamba-TransGhost model outperforms these benchmarks with a remarkable mAP@.5 of



a)

b)

Figure 2. Visualized results comparing the proposed model with the standard model in GDIT dataset.

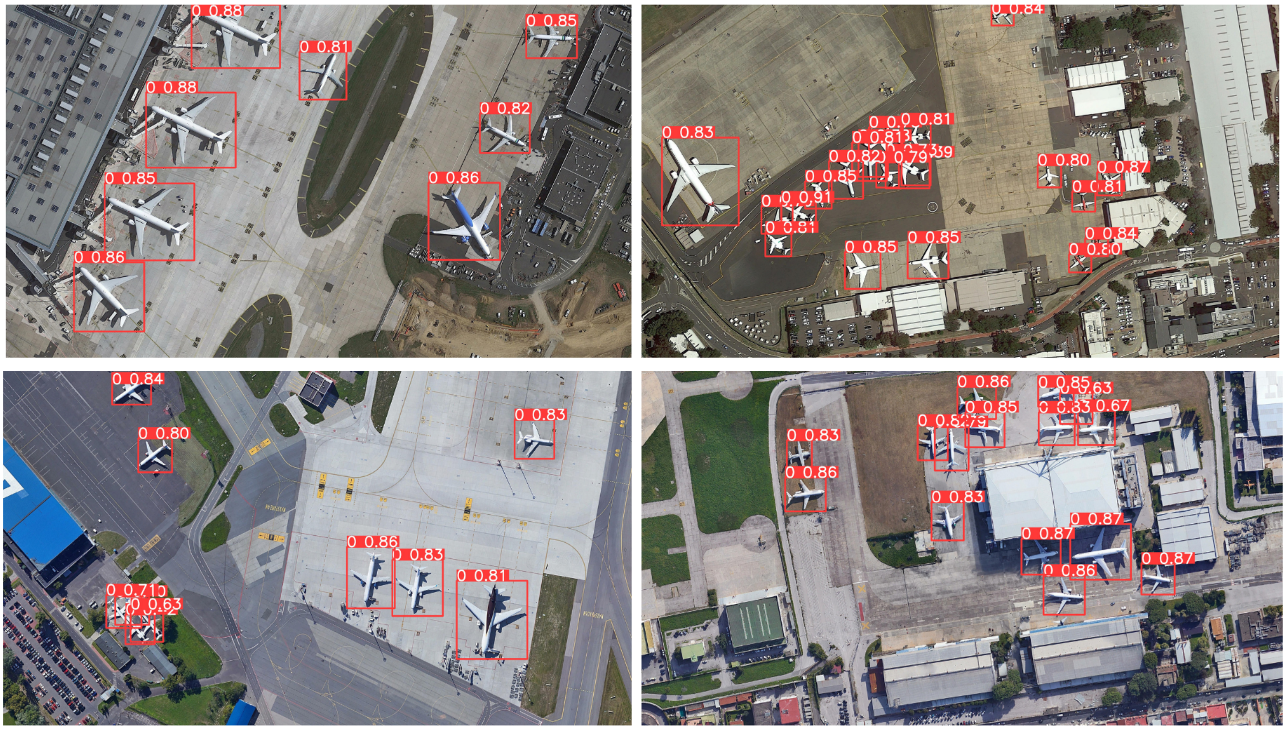


Figure 3. Detection results of the proposed model on the HRPlanev2 dataset.

94.30%, surpassing the highest-performing model from [22], YOLOv8, by 2.4%. Additionally, compared to the YOLO-NAS variants reported by [23], which achieved mAP@.5 scores of 81.50% and 81.80%, our model demonstrates a substantial improvement in detection accuracy.

In summary, the YOLO-Mamba-TransGhost model not only exceeds the accuracy of existing models but also sets a new standard in detection performance for satellite imagery analysis. This progress underscores the model's superior capability and relevance in real-time aircraft detection. Despite these advancements, challenges such as weather-related distortions and the need for larger, more diverse datasets remain. Future work will focus on further enhancing the model through advanced backbone networks and loss functions, and exploring real-time deployment and optimization strategies.

4.5. Visualization

The quantitative results provided earlier highlight the effectiveness of the YOLO-Mamba-TransGhost model. To visually demonstrate this, Figure 2 shows a comparison of detection results using the GDIT test dataset. The enhanced YOLO-Mamba-TransGhost model excels at detecting small-sized objects across a range of distances, addressing the detection gaps observed with the original YOLO-Mamba model and achieving notably improved accuracy.

Additionally, Figure 3 presents the visual results of the YOLO-Mamba-TransGhost model applied to the HRPlanev2 dataset. These images illustrate the model's capability to detect aircraft in various scenarios, including distant and close-up views, as well as in cases where aircraft colors are similar to their surroundings. This demonstrates the model's robustness and adaptability in diverse detection contexts.

5. CONCLUSIONS

In this study, we introduced and evaluated the YOLO-Mamba-TransGhost model for aircraft detection in satellite imagery, addressing challenges related to resolution, background complexity, and diverse object sizes and orientations. By refining the YOLO-Mamba architecture and incorporating the SC3T Transformer module along with the Ghost Convolution and WIoU loss function, we achieved notable advancements in detection performance and accuracy.

Our experiments on the HRPlanesV2 and GDIT datasets demonstrated significant improvements. The YOLO-Mamba-TransGhost model exhibited superior performance, with an mAP@.5 score of 94.30% on the GDIT dataset and 99.20% on the HRPlanesV2 dataset, reflecting its enhanced ability to detect and localize aircraft accurately under varying conditions. This advancement is achieved despite an increase in model complexity, highlighting the trade-off between accuracy and computational demands.

In comparison with other established models, our YOLO-Mamba-TransGhost model shows a clear advantage in detection accuracy, setting a new benchmark for aircraft detection in satellite imagery. Although challenges remain, such as handling weather-related distortions and the need for more diverse datasets, our findings underscore the model's potential in advancing remote sensing technologies.

Future work will focus on further optimizing the model through the integration of more advanced feature extraction networks and exploring real-time deployment scenarios. We aim to contribute to the broader field of computer vision by addressing current limitations and pushing the boundaries of satellite image analysis.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[2] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137-1149, 2016.

[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*, vol. 9905 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 21-37.

[5] Do, M. T., Ha, M. H., Nguyen, D. C., Tzyh-Chiang Chen, O. "Toward improving precision and complexity of transformer-based cost-sensitive learning models for plant disease detection," *Frontiers in Computer Science*, 6, 1480481, 2025.

[6] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[7] T. L. C. Tran, Z. C. Huang, K. H. Tseng, and P. H. Chou, "Detection of Bottle Marine Debris Using Unmanned Aerial Vehicles and Machine Learning Techniques," *Drones*, vol. 6, no. 12, p. 401, 2022.

[8] D. Xu and Y. Wu, "Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection," *Sensors*, vol. 20, no. 15, p. 4276, 2020.

[9] L. Zhou, H. Yan, Y. Shan, C. Zheng, Y. Liu, X. Zuo, and B. Qiao, "Aircraft detection for remote sensing images based on deep convolutional neural networks," *Journal of Electrical and Computer Engineering*, vol. 2021, no. 1, p. 4685644, 2021.

[10] Y. Yang, G. Xie, and Y. Qu, "Real-time detection of aircraft objects in remote sensing images based on improved YOLOv4," in *2021 IEEE 5th Advanced*

Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 1156-1164.

[11] Y. Kun, H. Man, and Y. L. Yanling, "Multi-target detection in airport scene based on YOLOv5," in *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 2021, pp. 1175-1177.

[12] Y. Li and X. Zhang, "Object detection for UAV images based on improved YOLOv6," *IAENG International Journal of Computer Science*, vol. 50, no. 2, pp. 759-768, 2023.

[13] M. T. Do, M. H. Ha, D. C. Nguyen, K. Thai, and Q. H. Do Ba, "Human Detection Based YOLO Backbones-Transformer in UAVs," in *2023 International Conference on System Science and Engineering (ICSSE)*, 2023, pp. 576-580.

[14] M. Bakirci and I. Bayraktar, "Transforming aircraft detection through LEO satellite imagery and YOLOv9 for improved aviation safety," in *2024 26th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2024, pp. 1-6.

[15] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[16] Z. Wang, C. Li, H. Xu, and X. Zhu, "Mamba YOLO: SSMs-Based YOLO For Object Detection," *arXiv preprint arXiv:2406.05835*, 2024.

[17] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: bounding box regression loss with dynamic focusing mechanism," *arXiv preprint arXiv:2301.10051*, 2023.

[18] M. T. Do, T. D. Kim, M. H. Ha, O. T. C. Chen, D. C. Nguyen, and A. L. Q. Tran, "An Effective Method for Detecting Personal Protective Equipment at Real Construction Sites Using the Improved YOLOv5s with SIoU Loss Function," in *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2023, pp. 430-434.

[19] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580-1589.

[20] S. E. Ghazouali, A. Gucciardi, N. Venturi, M. Rueegsegger, and U. Michelucci, "FlightScope: A Deep Comprehensive Assessment of Aircraft Detection Algorithms in Satellite Imagery," *arXiv preprint arXiv:2404.02877*, 2024.

[21] T. Bakirman and E. Sertel, "A benchmark dataset for deep learning-based airplane detection: HRPlanes," *arXiv preprint arXiv:2204.10959*, 2022.

[22] G. H. Rishi, P. R. Kumar, N. C. Varshit, and K. Sailaja, "Real-Time Military Aircraft Detection using YOLOv5," in *2024 Second International Conference on Data Science and Information System (ICDSIS)*, 2024, pp. 1-7.

[23] Medium, "YOLO-NAS: A game-changer in object detection with Deci AI's neural architecture search technology," *Medium*, 2024.