scientific reports

Check for updates

OPEN

A novel hybrid deep learning approach for super-resolution and objects detection in remote sensing

Muhammad Asif[1], Mohammad Abrar[2], Faizan Ullah[1], Abdu Salam[3], Farhan Amin[4]✉, Isabel de la Torre[5]✉, Mónica Gracia Villar[6], Helena Garay[6] & Gyu Sang Choi[4]✉

Object detection in remote sensing imagery presents challenges due to low resolution, complex backgrounds, occlusions, and scale variations, which are critical in disaster response, environmental monitoring, and surveillance. This study proposes a robust object detection framework integrating super-resolution techniques with advanced feature extraction algorithms for remote sensing images. The hybrid model combines Advanced StyleGAN and Swin Transformer. Advanced StyleGAN enhances image resolution, facilitating the detection of small and occluded objects, while Swin Transformer employs hierarchical attention mechanisms for effective feature extraction. Preprocessing techniques, including data augmentation, are incorporated to improve the diversity and accuracy of the training dataset. Evaluation on datasets such as VEDAI-VISIBLE and VEDAI-IR demonstrated exceptional performance, achieving an mAP@0.5 of 97.2%, mAP@0.5:0.95 of 72.8%, and F1-Score of 0.93, with an inference time of 42 ms. The framework maintained robustness under challenging conditions, such as low light and fog, outperforming YOLOv9-S, YOLOv9-E, and DCNN-based methods. Furthermore, it surpassed state-of-the-art models on RSOD and NWPU VHR-10 datasets, achieving superior detection accuracy and robustness. This framework offers a significant advancement in remote sensing object detection, providing an effective solution for complex scenarios. Future work may focus on optimizing computational efficiency and expanding the framework to multimodal or dynamic object detection tasks.

Keywords Super-resolution, High resolution, Object detection, Deep learning, Remote sensing

With the recent rapid expansion of remote sensing enterprises, quick and efficient access to road information from remote sensingimages is becoming possible. With that, remote sensing has become a significant area of research and has been applied in a wide range of areas, such as environmental conservation, urban development, disaster response, and military intelligence in the past decade[1]. Technological advancements have allowed the acquisition of more detailed images from aerial and satellite sources that can be used to analyze better and interpret vast quantities of data from the Earth's surface. Although such advancements have been made, the identification of small objects present within low-resolution remotely sensed images remains a persistent challenge[2]. Small objects are challenging to detect and classify due to factors such as sensor limitations, cluttered backgrounds, and varying environmental conditions[3]. Super-resolution techniques have been the topic of research to improve low-resolution images for object detection purposes by reconstructing high-resolution images. Earlier super-resolution methods utilized interpolation techniques, which resulted in blurred and lackluster images. Unfortunately, however, until very recently, deep learning had not progressed far enough within remote sensing imagery to overcome the noise in the imagery and bring about meaningful results, generating images that had not yet been super-resolved and did not reveal much more detail in objects of interest in the imagery[4].

Despite this, the problem of detecting small objects in low-resolution images remains highly desired. Although many studies have been done on this problem, GAN-based super-resolution models, including ESRGAN, still face difficulties preserving fine details and reaching the best object detection results in complex remote sensing environments[5]. These performance gaps indicate a need for more investigation of more sophisticated GAN

[1]Department of Computer Science, Bacha Khan University, Charsadda 24420, Pakistan. [2]Faculty of Computer Studies, Arab Open University, Muscat 122, RiyadhP.O. Box 1596, Oman. [3]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan. [4]School of Computer Science and Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea. [5]Department of Signal Theory and Communications, University of Valladolid, Valladolid, Spain. [6]Universidad Europea del Atlántico, Santander, Spain. ✉email: farhanamin10@hotmail.com; isator@uva.es; castchoi@yu.ac.kr

structures, like StyleGAN, that have shown outstanding performance in generating photorealistic images with high realism and fine details in other domains.

For a long time, researchers have been interested in identifying small objects in low-resolution images. Yet, to address this problem, imaging techniques have been developed, and deep learning-based methods such as ESRGAN have been developed. Still, they do not render object detection accurate enough, especially in complex cases. In addition, CNN-based architectures, employed for object tracking, are constrained by their fixed receptive fields and local operations that prevent using CNNs for computing long-range dependencies and the global context to find small, occluded, or scale variant objects in remote sensing systems. This paper proposes using Advanced StyleGAN for super-resolution remote sensing images to overcome these challenges. It is proposed to use the Advanced StyleGAN model to improve image resolution to enable the detection of small objects which would otherwise be difficult to identify[6,7]. Furthermore, this paper attempts to deal with the shortcomings of super-resolution and object detection in remote sensing by integrating Advanced StyleGAN with Swin Transformer to enhance object detection in low-resolution remote sensing images. Both technologies' strengths are combined in the proposed approach to address the problems in remote sensing while also opening a new door for more efficient analysis of satellite and aerial images.

The research presents a new framework combining Advanced StyleGAN with Swin Transformer to achieve super-resolution and object detection. The proposed method substantially improves the detection of small targets in low-resolution remote sensing images. The procedure boosts image features and detail definition to provide better performance than existing systems. Moreover, the benchmark dataset analysis confirms that the integrated framework is efficient and effective and establishes itself as a new benchmark for remote sensing object detection.

The primary contributions of this study are as follows:

- Novel Integration: Introducing a unified framework that combines Advanced StyleGAN for super-resolution with the Swin Transformer for object detection, bridging the gap between GAN-based and transformer-based models in remote sensing.
- Improved Object Recognition: This effectively addresses the challenges of detecting small, occluded, and scale-invariant objects in low-resolution images, surpassing existing methods in accuracy and computational efficiency.
- Robust Generalization: Validating the proposed approach on varied remote sensing datasets, VEDAI VISIBLE, VEDAI IR, and DOTA, demonstrating its adaptability to complex scenes and environmental conditions.
- Hierarchical Attention Mechanisms: Incorporating advanced attention mechanisms in the Swin Transformer to enhance detection performance regardless of object size or scene complexity.

Hardware constraints and resource limitations are crucial in system design in many real-world remote sensing scenarios. Although our primary goal is to improve detection accuracy for small or occluded objects, this approach remains mindful of the trade-off between computational overhead and performance. Acknowledging these limitations early on, it aims to develop a strategy that can eventually be adapted to edge devices, mobile platforms, or other resource-constrained environments where memory usage, inference speed, and power consumption are critical.

The rest of this paper is structured as follows: Sect. 2 reviews existing super-resolution techniques, GAN-based image enhancement, and object detection models. Section 3 details the proposed methodology, datasets, Advanced StyleGAN for super-resolution, Swin Transformer for object detection, and evaluation metrics. Section 4 presents experimental results, a comparative analysis of the proposed and existing approaches, and performance evaluations. Finally, Sect. 5 concludes the paper and suggests directions for future research.

Related work

This section reviews recent advancements in super-resolution methods, GANs for image enhancement, and object detection frameworks in remote sensing. It highlights key studies that have significantly influenced current approaches.

Super-Resolution techniques in remote sensing

Remote sensing greatly relies on super-resolution (SR) performance to increase the spatial resolution of the satellite data and contribute to a more accurate analysis. The needed application spaces for this advancement include environmental monitoring, urban planning, and disaster management systems. The SpectralGPT model[8] represents a significant development because it uses a 3D generative pre-trained transformer designed for spectral remote sensing images. This model demonstrates exceptional effectiveness in reducing errors in scene recognition and changing identification tasks, which has the potential to boost numerous remote sensing tasks. A fundamental exploration of remote sensing resolution and scale effects through enhanced pixel resolution is presented, which serves as a foundation for SR research[9]. The RingMo framework serves as a self-supervised learning model that generates high-resolution remote-sensing images, according to[10]. RingMo demonstrates the highest level of performance across multiple tasks through its analysis of extensive datasets, which opens potential improvements for self-supervised SR learning in applications. Remote sensing object detection receives a boost from the large selective kernel network (LSKNet)[11] using an optimized spatial receptive field to achieve better detection results. Modern scientific developments show that remote sensing data can be improved through a synergistic combination of SR techniques for enhanced image analysis capabilities.

Generative adversarial networks (GANs) for image enhancement

GANs have become essential for image enhancement, especially in remote sensing, where high-resolution imagery is essential for detailed analysis. Due to its excellent performance in transforming low-quality images to high-resolution images, GAN benefits several applications, including object detection, land cover classification, and environmental monitoring. The integration of GANs with better Mask R-CNN models can be applied in one application: improving edge detection in satellite images, where increased accuracy is required for monitoring disturbed areas in construction[12]. In change detection models, GANs have also been employed in dual-branch multilevel intertemporal networks (DMINet) for effectively capturing small changes between bitemporal images that are important for monitoring dynamic environments[13]. Additionally, GANs have demonstrated utility in enhancing facial image resolution while retaining intrinsic characteristics[14]. Moreover, a review article concurs that GAN-based models surpass the traditional deep learning counterpart that problem needs a lot of detail and accuracy[15]. In meteorology and oceanography, GANs have been used to reconstruct missing information on a small scale in turbulent flow fields using lower-resolution images[16]. Further, PROBA-V satellite imagery is enhanced with GAN-based super-resolution methods to increase the resolution and detail of the vegetation monitoring[17]. Additionally, GAN was employed to increase the resolution of hyperspectral images for environmental mapping as well as agriculture[18]. In addition, GAN-based models have been employed to upscale urban surveillance images for planning and security applications[19].

Object detection models in remote sensing

Object detection is an important problem in remote sensing and can be utilized in land use mapping, urban planning, disaster management, and environmental monitoring. Recent advances in deep learning have dramatically increased the capacity to detect objects, buildings, vehicles, and natural features in satellite and aerial imagery. Faster R-CNN is among the most popular models because it uses a region proposal network to locate the regions of interest. The model has been used to detect objects such as ships, aircraft, or vehicles in high-resolution satellite images[20]. The combination of Faster R-CNN with feature pyramid networks (FPN) has further boosted object detection, especially for small objects (occurring in remote sensing due to variances in object size from scene to scene), better than what can be achieved via a single-stage detector. Real-time results are also being yielded through You Only Look Once (YOLO) in remote sensing object detection. The high accuracy and speed provided by YOLO models, such as the latest versions, YOLOv5 and YOLOv7, make them suitable for large-scale image analysis in remote sensing[21]. The single-shot Multibox detector (SSD) has become very popular, especially when dealing with real time applications like traffic and city surveillance[22]. To deal with class imbalance problems in object detection, RetinaNet utilizes focal loss that helps to detect small and densely occluded objects with higher accuracy while preserving the detection accuracy of larger objects[23]. Among the applications of the Mask R-CNN model, including building footprint extraction and disaster damage assessment[24], its outstanding object detection and instance segmentation ability has been proved. Furthermore, transformer-based models like detection transformer (DETR) have gotten rid of region proposals using attention mechanisms and improved object detection in such complex scenes[25].

Materials and methods

This section describes the proposed hybrid model integrating Advanced StyleGAN for super-resolution and the Swin Transformer for object detection. It also details the data preprocessing steps, model training, tuning, and evaluation metrics.

Proposed approach

The proposed approach utilizes Advanced StyleGAN to upscale low-resolution images to high-resolution ones, followed by the Swin Transformer for object detection. Advanced StyleGAN first takes low-resolution remote sensing images and generates their higher-resolution counterparts, preserving fine-grained details that might otherwise be lost. These super-resolved images then serve as input for our Swin Transformer-based detector. The benefit is twofold: (1) improved clarity for small or partially hidden objects, and (2) enhanced feature extraction for the Transformer. Swin Transformer achieves more precise object localization and classification based on how well-detailed an image is. Super-resolution techniques working together with attention-based detection methods enhance identification accuracy when dealing with various levels of scene disarray. An image super-resolution system functions through three essential parts: the generator discriminator and specific loss functions to enhance output image authenticity.

The generator begins with low-resolution images to generate high-resolution images on its output side. The model produces artificial images before a classifier evaluates real high-definition images and the fake images it produces. A perceptual loss function enhances the quality of output images to produce super-resolved images that appear sharp with defined details and a textured appearance.

The image outputs produced by Advanced StyleGAN become the input of the Swin Transformer, which specializes in remote sensing applications. Image enhancement occurs through Advanced StyleGAN, yet the Swin Transformer enhances accuracy and operational efficiency when detecting small objects in challenging remote sensing conditions. The core function of Advanced StyleGAN involves creating high-resolution images from low-resolution data that keeps essential textural patterns intact. These enhanced images are then fed directly into the Swin Transformer detector. Because the Swin Transformer excels at extracting multi-scale features through hierarchical attention, having sharper, more detailed inputs significantly boosts its ability to localize and classify small or partly occluded objects. This tight combination of super-resolution and attention-based detection underpins our system's high accuracy, especially in complex remote sensing conditions. The entire methodology is visually summarized in Fig. 1 for clarity.

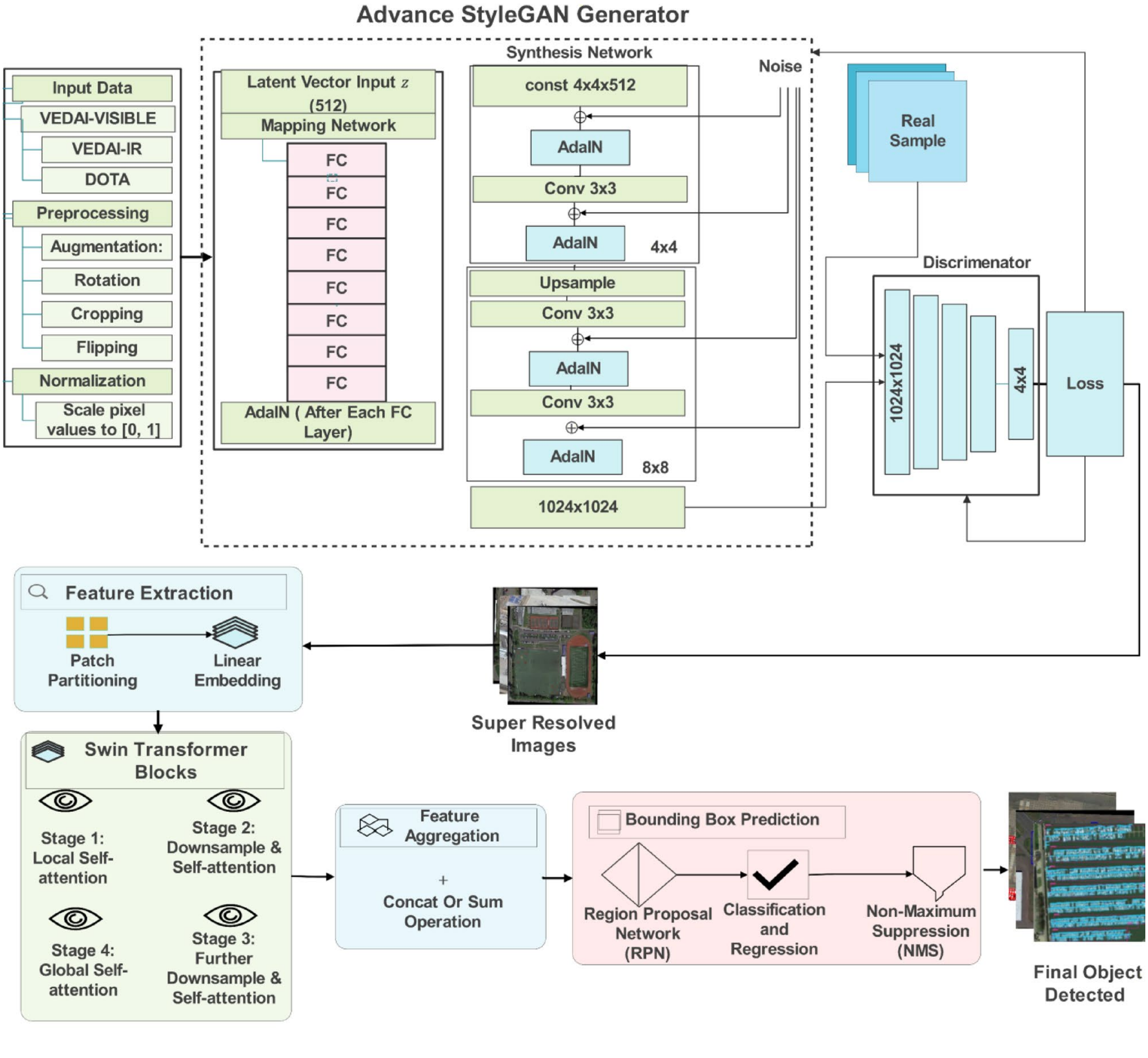
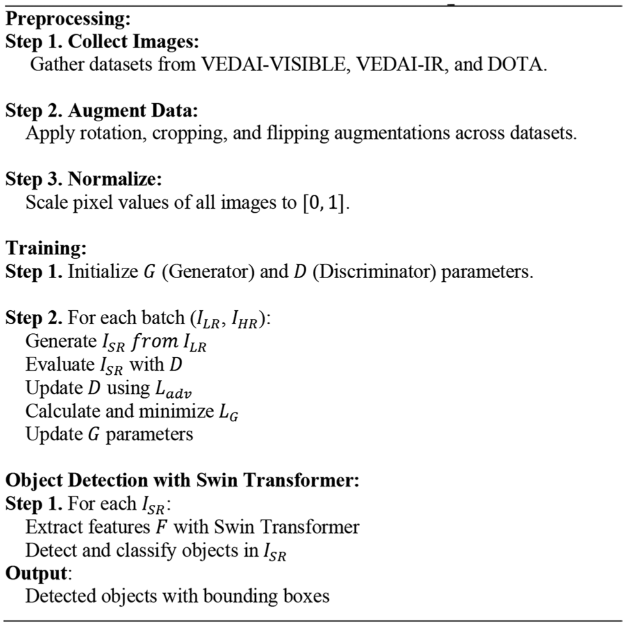


Fig. 1. Proposed Framework.

As shown in Fig. 1, the proposed framework integrates Advanced StyleGAN and Swin Transformer to address the challenges in object detection in remote sensing imagery. Data from the remote sensing datasets VEDAI_VISIBLE, VEDAI_IR, and DOTA form the input. The input data is prepared for subsequent stages through preprocessing steps such as rotation, cropping, flipping, and normalization. The Advanced StyleGAN generator processes the preprocessed data to transform it into high-resolution images using a mapping network and adaptive instance normalization (AdaIN) layers. To achieve this, the generator is trained with a discriminator, and the loss is minimized to keep the generated images having enhanced detail and texture. Then, these super-resolved images with more information are fed into the Swin Transformer for feature extraction.

The Swin Transformer's hierarchical attention framework enables high-resolution feature learning while it processes images from beginning to end to receive broad contextual understanding later in processing. Through this process, the model achieves better resistance to object size differences and improves its ability to handle obstructed and varied scale targets.

The region proposal network (RPN) produces bounding boxes from all extracted features through classification and regression, further applying non-maximum suppression. The combined framework of Advanced StyleGAN with Swin Transformer generates a solution that provides precise object detection performance on remote sensing imagery containing small or partially obscured complex objects.



```
Preprocessing:
Step 1. Collect Images:
    Gather datasets from VEDAI-VISIBLE, VEDAI-IR, and DOTA.

Step 2. Augment Data:
    Apply rotation, cropping, and flipping augmentations across datasets.

Step 3. Normalize:
    Scale pixel values of all images to [0, 1].

Training:
Step 1. Initialize $G$ (Generator) and $D$ (Discriminator) parameters.

Step 2. For each batch ($I_{LR}$, $I_{HR}$):
    Generate $I_{SR}$ $from$ $I_{LR}$
    Evaluate $I_{SR}$ with $D$
    Update $D$ using $L_{adv}$
    Calculate and minimize $L_G$
    Update $G$ parameters

Object Detection with Swin Transformer:
Step 1. For each $I_{SR}$:
    Extract features $F$ with Swin Transformer
    Detect and classify objects in $I_{SR}$
Output:
    Detected objects with bounding boxes
```

Algorithm 1. Advanced StyleGAN-Swin transformer for super-resolution and object detection.

Datasets

Researchers applied careful preprocessing techniques to their datasets to contain image variations alongside structural differences between the image sets. This study utilized the VEDAI-VISIBLE database[26], and VEDAI-IR database[27]. Remote sensing image detection of small objects forms the primary focus of VEDAI-VISIBLE, but DOTA[28] increases the examination complexity for better detection accuracy. The chosen datasets represent different usual conditions and complexities in remote sensing applications. VEDAI provides a closer look at small vehicles (visible and infrared), DOTA features large-scale overhead images with complex backgrounds, and RSOD/NWPU VHR-10 includes diverse classes, object orientations, and resolutions. Combining datasets with different scales and spectral bands ensures that our model learns to detect objects under many types of clutter, occlusion, and environmental conditions. This selection also reflects widely recognized benchmarks in the field, allowing us to compare our approach to other published work more directly.

The VEDAI datasets contain 1,210 images, each withdimensions of 1024 × 1024 pixels in the visible and infrared spectrums. Structural complexity in these datasets was defined based on the number of classes per image: images with fewer classes were categorized as having low complexity, whereas those with more classes exhibited moderate complexity.

On the other hand, the DOTA dataset is larger and more diverse, consisting of 2,806 images ranging from 800 × 800 to 20,000 × 20,000 pixels. Structural complexity was categorized similarly, with images containing fewer than two classes labeled as low complexity, those with three to four classes as moderate complexity, and images with more than four classes classified as highly complex.

The images were divided into four sets of increasing complexity and size to prepare the datasets systematically. This progression ensured the gradual introduction of structural diversity and variability.

- **Set 1**: Derived from the VEDAI-VISIBLE dataset, this set included 10% of the dataset, comprising 121 images labeled manually into two classes: cars and trucks. Structural complexity was low, and no image augmentation was applied.
- **Set 2**: This set builds upon Set 1 and contains images from the VEDAI-IR dataset. Rotations of 90°, 180°, and 270° were performed on image augmentation, resulting in 484 images for the dataset. Structural complexity was moderate.
- **Set 3**: This set consisted of images from VEDAI-VISIBLE and VEDAI-IR, combined with augmentations such as cropping and horizontal and vertical flipping. The dataset increased from the previous sets of 207 to 726, and the structural complexity increased.
- **Set 4**: The images from previous sets combined with the data of others from the DOTA dataset. The same augmentations, such as rotations and flips, were also applied to the new DOTA images. This set had 1,932 high complex structures, including diverse objects and conditions.

By preparing this structured dataset, the model could train on increasingly complex images beyond the variety it would encounter in real-world scenarios. This approach strengthened the model's generalization and remote sensing detection.

Advanced stylegan for super-resolution

This high-performance generative model enhances the quality of remotely sensed images to detect small objects more easily. Building on the foundation of StyleGAN has been demonstrated to produce high-quality image synthesis with the ability to manipulate different levels of detail; this model extends the foundation. Specific to remote sensing imagery, this implementation makes modifications and uses the properties of remote sensing imagery to achieve the best performance for such data.

Architecture and design

The two core components of the Advanced StyleGAN are the generator G and the discriminator D.

The generator aims at mapping a low-resolution image I_{LR} to a high-resolution image I_{SR}. It comprises several convolutional layers, AdaIN layers to align the input image's statistics, and skip connections to preserve image details. In mathematical terms, the generator's output is:

$$I_{SR} = G(\ I_{LR}, z) \tag{1}$$

Expressly, it is conditioned on z being a latent vector sampled from a Gaussian distribution fed through style modulation layers at different generator stages. It also allows for controlling fine details in the output image to varying levels of detail, which is particularly useful for enhancing remote sensing features.

The discriminator evaluates the authenticity of the generated high-resolution images, computing a probability score to determine if the images are real high-resolution images I_{HR} or not. The discriminator is trained to minimize the adversarial loss:

$$L_{adv} = E_{I_{LR}}\left(\left[\log D\ \left(I_{LR}\right)\right] +\ E_{I_{LR}}\left(\left[log - 1D\ \left(G\left(I_{LR}, z\right)\right)\right)\right]\right. \tag{2}$$

Training process and loss functions

The training process for Advanced StyleGAN involves iterative optimization of the generator and discriminator using a combination of loss functions designed to balance image fidelity and realism.

- Adversarial loss (L_{adv}): Ensures the generator produces images indistinguishable from real high-resolution images.
- Perceptual loss ($L_{perceptual}$): Ensures that the generated images retain perceptual similarity to high-resolution images. This is computed by comparing feature maps extracted from a pre-trained VGG network:

$$L_{perceptual} = \sum{}_{i=1}^{N} \frac{1}{C_i \times\ H_i \times\ W_i}\ \|\ \varnothing_{\,i}(I_{LR}) - \varnothing_{\,i}(I_{SR})\|^{\,2} \tag{3}$$

where $\varnothing_{\,i}$ represents the feature maps from the i^{th} layer of the Swin Transformer, and C_i, H_i, and H_i denote the channels, height, and width, respectively.

- Content loss ($L_{content}$): Preserves the structural content of the original low-resolution image:

$$L_{content} = \|\ I_{LR} - I_{SR}\|^{\,2} \tag{4}$$

- Style loss (L_{style}): Captures the style of the original image, including texture and color distribution, using the Gram matrix of feature maps:

$$L_{style} \sum{}_{i=1}^{N} \|\ G\left(\varnothing_{\,i}\left(I_{HR}\right)\right) - G\left(\varnothing_{\,i}\left(I_{SR}\right)\right)\|^{\,2} \tag{5}$$

The overall loss function for the generator is a weighted sum of these losses:

$$L_G = \lambda_{\,adv} L_{adv} +\ \lambda_{\,perceptual} L_{perceptual} + \lambda_{\,content} L_{content} + \lambda_{\,style} L_{style} \tag{6}$$

where $\lambda_{\,adv}$, $\lambda_{\,perceptual}$, $\lambda_{\,content}$, and $\lambda_{\,style}$ are weights controlling the contribution of each loss.

The training process enhances image resolution and facilitates robust object detection in remote sensing applications. The systematic methodology for this process is detailed in Algorithm 1, which outlines the step-by-step approach for achieving these objectives.

Swim transformer for object detection

The Swin Transformer is utilized for object detection by leveraging its hierarchical feature extraction and window-based attention mechanisms. This approach allows for accurate detection of objects in super-resolved remote sensing images, as it captures both local and global contexts. The images are processed through the model, which extracts multi-scale features, enabling the detection of objects across different sizes. The final output provides bounding boxes and class probabilities, ensuring high accuracy and fast processing, even in complex RS scenarios.

Architecture of the Swin transformer

For object detection, the Swin Transformer takes the super-resolved image I_{SR}, generated by the Advanced StyleGAN, as input. The image is first split into non-overlapping patches of size P × P, which are then embedded into a sequence of patch embeddings X_0. These embeddings serve as the input to the Swin Transformer:

$$X_0 = \text{PatchEmbed}\left(I_{SR}\right) \tag{7}$$

Each stage of the Swin Transformer processes these patch embeddings through a combination of the Multilayer Perceptron (MLP) and Shifted Window Multi-Head Self-Attention (SW-MSA) layers. The transformation at the layer l can be represented as:

$$X_{i=1} = MLP(SW - MSA\left(X_l\right) + \left(X_l\right) + X_l \tag{8}$$

where X_i represents the features at the layer l, and the attention mechanism SW-MSA is performed within the shifted window. This allows the model to learn dependencies across different regions of the image. The MLP further refines these features, enhancing the model's ability to identify objects at varying scales.

During model training, patch merging occurs at multiple stages to reduce spatial dimensions and expand the receptive field, which is particularly useful for detecting objects of varying sizes:

$$X_{down} = PatchMerge\left(X_{prev}\right) \tag{9}$$

where $PatchMerge$ combines adjacent patches to create a coarser, more abstract feature representation.

Integration with Super-Resolved images

For object detection, super-resolved images are generated using the Advanced StyleGAN and fed directly into the Swin Transformer to extract features. The integration of super-resolution and object detection using the Swin Transformer can be described mathematically as follows:

The low-resolution image I_{LR} is transformed into a super-resolved image I_{SR} using the Advanced StyleGAN (Eq. 1).

The super-resolved image I_{SR} is then passed into the Swin Transformer, where its hierarchical attention-based architecture is employed to extract features and predict object locations and class labels:

$$Y_{output} = H_{Swin}\left(I_{SR}\right) \tag{10}$$

where H_{Swin} represents the Swin Transformer model, and Y_{output} includes the predicted bounding boxes and class labels.

This approach leverages the fine detail in the super-resolved images, allowing Swin Transformer to accurately discover and categorize objects even in difficult remote sensing scenes where the object may be small or partially obscured.

Training and fine-tuning strategies

The Swin Transformer, alongside Advanced StyleGAN, utilized NVIDIA RTX 3090 GPUs with PyTorch as their basis to conduct their training. Training and fine-tuning represented the two principal phases of the learning procedure.

The Advanced StyleGAN received training on VEDAI-VISIBLE and VEDAI-IR images to create detailed images starting from basic input resolutions. Adam optimizer trained the process with a learning rate set at 0.0001 while conducting pre-training procedures. The training utilized 16 items per batch, decreasing the learning rate by cosine anneal functions during 100 epochs. The models proceeded with training following the super-resolution model and achieved adequate competence where high-resolution images served to train the Swin Transformer for object detection. The Swin Transformer fine-tuned its pre-trained weights while an additional reduction was applied to the learning rate to counteract overfitting. Random rotation flipping and scaling served as data augmentation approaches during this process. Table 1 provides the complete set of hyperparameters for both Advanced StyleGAN and Swin Transformer.

Hyperparameter	Advanced StyleGAN	Swin Transformer
Optimizer	Adam	AdamW
Learning Rate	0.0001	0.00005
Learning Rate Schedule	Cosine Annealing (100 epochs)	Decay Factor: 0.9 (every 10 epochs)
Batch Size	16	8
Training Epochs	100	150
Loss Functions	Adversarial, Perceptual, Content, Style	Focal Loss, GIoU Loss
Data Augmentation	Rotation, Flipping	Rotation, Flipping, Cropping
Input Image Size	N/A	512 × 512 pixels
Pre-trained Weights	N/A	Pre-trained on the COCO dataset
Hardware	NVIDIA RTX 3090 (24 GB VRAM)	NVIDIA RTX 3090 (24 GB VRAM)
Framework	PyTorch	PyTorch

Table 1. Training hyperparameters for advanced stylegan and Swin transformer.

Metric	Formula
Mean Average Precision (mAP)	$mAP = \frac{1}{\lvert C\rvert}\sum_{c\in C} AP\left(c\right)$
F1-Score	$F1 = 2.\frac{Precision.Recall}{Precision+Recall}$
Precision	$Precision = \frac{TP}{TP+FP}$
Recall	$Recall = \frac{TP}{TP+FN}$
Peak Signal-to-Noise Ratio (PSNR)	$PSNR = 10.log_{10}\left(\frac{L^2}{MSE}\right)$, where L is the maximum possible pixel value and MSE is the mean squared error.
Structural Similarity Index (SSIM)	$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+c_2)}$, where μ, σ, and C_1, C_2 are statistical properties and constants.
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$, where y_i is the ground truth value and $\widehat{y_i}$ is the predicted value.
Inference Time	Measured directly in milliseconds (ms) using system timers during inference.
GPU Utilization	Directly monitored using GPU profiling tools (e.g., NVIDIA's GPU Profiler).
Memory Usage	Measured in gigabytes (GB) using profiling tools.

Table 2. Formulas for evaluation metrics used in the proposed approach.

Dataset	PSNR (dB)	SSIM	MSE
VEDAI-VISIBLE	30.2	0.92	0.0018
VEDAI-IR	29.8	0.91	0.0020
DOTA	28.5	0.89	0.0025

Table 3. PSNR, SSIM, and MSE scores for advanced StyleGAN.

Evaluation metrics

Several performance metrics are used to evaluate the proposed model's performance. The first was the mean average precision (mAP) concerning different Intersections over union (IoU) thresholds, which was the object detection precision for a set of categories. The F1 score was also used to measure the model's object detection, classification accuracy, and recall. The model was also evaluated regarding its feasibility in remote sensing applications and the evaluation of inference time and computational resource usage. All evaluation metric formulas are given in Table 2.

Experiments and results

An evaluation of the described Advanced StyleGAN super-resolution with Swin Transformer object detection method in RS images uses the steps outlined in this section. The review analyzed three performance metrics from the model, including its precision in detection tasks, computational speed during operation, and ability to handle complex RS application contexts.

Implementation details and experimental setup

The experimental design ran on PyTorch deep learning programming software and was executed on NVIDIA RTX 3090 GPU systems. The VEDAI-VISIBLE, VEDAI-IR, and DOTA datasets provided diverse environmental conditions for the training and evaluation phases. All images received 512 × 512-pixel resolution for standardization between datasets.

Performance of advanced stylegan

The advanced StyleGAN system demonstrated its capacity for super-resolving images through objective quality evaluation using structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and mean squared error (MSE). The metrics evaluate both the SR image quality and verify the authenticity of the ground truth. Table 3 shows the performance results of objective quality metrics from the VEDAI-VISIBLE, VEDAI-IR, and DOTA datasets.

The Advanced StyleGAN maintained high PSNR and SSIM scores in all datasets because it showed excellence in resolution enhancement and detail preservation. The model demonstrates strong error precision between super-resolution output and reference ground truth by producing small MSE values. The capability of Advanced StyleGAN to create high-quality SR images has been confirmed by these research findings, making them valuable for object detection operations.

The second evaluation consisted of a qualitative analysis that examined how Advanced StyleGAN processed super-resolved images to determine its success in preserving vital details in RS images. The analysis presented in Fig. 2 demonstrates database-to-database comparisons and shows how the proposed approach generates new results. The images in section (a) display low-resolution (LR), super-resolved (SR), and ground truth (GT) images before applying the method.

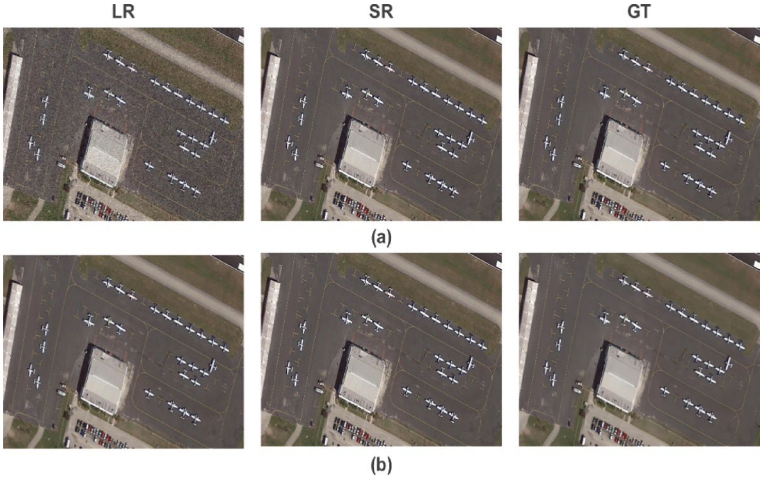


Fig. 2. Image quality before and after applying the proposed super-resolution method.

Dataset	mAP@0.5 Pre-Super-Resolution	mAP@0.5 Post-Super-Resolution	Improvement (%)
VEDAI-VISIBLE	72.5%	82.3%	+ 13.5%
VEDAI-IR	70.1%	78.6%	+ 12.1%
DOTA	65.4%	75.2%	+ 15.0%

Table 4. Object detection results pre- and post-super-resolution.

As shown in Fig. 2 (b), the super-resolution method proposed in this work significantly improves the image realism and detail compared to the original ones shown in Fig. 2 (a). The super-resolved images demonstrate improved image quality for RS applications. These enhanced images are particularly useful for precise object detection tasks where visual clarity is vital.

This research also evaluates the impact of super-resolution on object detection performance. Advanced StyleGAN is designed to improve the spatial resolution and realism of RS images, improving object detection models, particularly for small or occluded objects.

For all datasets, Table 4 shows consistent improvements in mAP at 0.5 IoU threshold after super-resolution, with gains between 12.1% and 15.0%. For instance, it increased the mAP from 72.5 to 82.3% for the VEDAI-VISIBLE dataset and had similar improvements for the VEDAI-IR and DOTA datasets. The super resolution process enhances the image quality, and it helps the detection model, particularly when combined with the Swin Transformer, to detect and classify more objects effectively. These results indicate that image quality is a key factor in effective remote sensing services, and super-resolution integration methods can dramatically improve the object detection process.

With the Swin Transformer, the detection model can better detect and classify objects in the RS images, making the use of super-resolution even higher. A hierarchical attention mechanism is used in the Swin Transformer to process super-resolved images so that the model can better localize object locations and categories. The overall feature enhancements in object detection performance are illustrated in this workflow (Fig. 3).

This approach integrated the super resolution with Swin Transformer, and because of the increased resolution and processed image quality, this achieved significant improvements in object detection. This combined approach is proven to be accurate, and the accuracy of the definition and classification of objects is highly dependent on the efficiency of this approach, which is of high value for complex scenarios where accurate identification is required.

Object detection accuracy with Swin transformer

MAP and F1-Score evaluation

Two metrics, mAP, and F1-Score, evaluate the Swin Transformer's object detection performance. These metrics help understand the model's precision and recall capability across different object categories. Table 5 shows the mAP values at two different IoU thresholds (0.5, 0.5:0.95). It shows that, in super-resolved RS images, the model achieves high efficiency in detecting multiple objects of various types.

The Swin Transformer has strong mAP values on all object categories. It performs well at the mAP@0.5 threshold, meaning the method can detect and classify objects in remote sensing images well. Finally, the mAP@0.5:0.95 values show that the model is precise, especially at higher IoU values (higher IoU means more overlap between the predicted bounding box and actual ground truth). Additionally, the F1 scores for each object category show that the proposed model combines precision and recall to detect many object types robustly. This

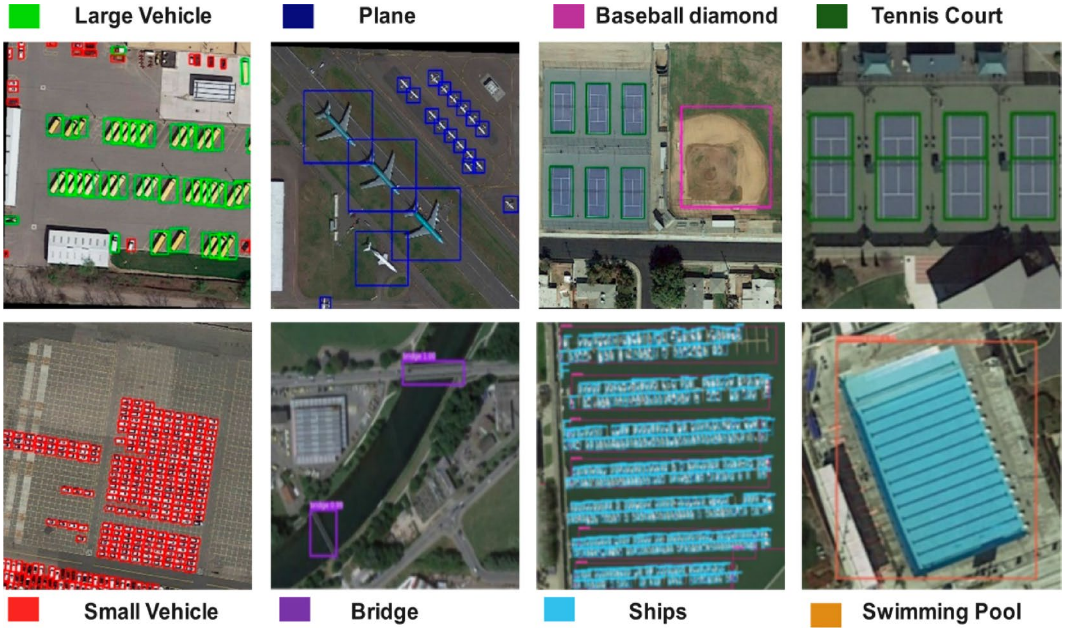


Fig. 3. Enhanced object detection through super-resolution and swin transformer.

Object category	mAP@0.5	mAP@0.5:0.95	F1-Score
Cars	85.4%	62.3%	0.88
Trucks	83.2%	60.1%	0.86
Planes	89.5%	68.7%	0.91
Ships	81.7%	59.8%	0.84
Storage Tanks	80.9%	57.4%	0.83

Table 5. mAP@0.5 and mAP@0.5:0.95 for various object categories.

corroborates the fact that the Swin Transformer is a good fit for the problem of remote sensing imagery when combined with super-resolved inputs from Advanced StyleGAN.

For the three key datasets, VEDAI-VISIBLE, VEDAI-IR, and DOTA, the Swin Transformer has been experimented with in the context of object detection and scrutinized concerning its precisions and recalls. These metrics were used for training to check how well the model could distinguish true and false positives. Figure 4 shows the accuracy and recall curves for each model dataset. The Swin Transformer eventually gets more precise and recalls all three datasets. Finally, looking at the precision curves, the curves get more accurate when reducing the false positives, whereas, on the recall curves, The curves also have a better ability to find the true positives (especially in the case of DOTA). It demonstrates the Swin Transformer's versatility and high accuracy and can strike a good balance between precision and recall in different remote sensing scenarios. Overall, the metrics are improving continuously, which means that the model is ready and can manage the complexities of object detection in a complicated environment.

Confusion matrix analysis

The confusion matrix is used to evaluate the effectiveness of the object detection of the Swin transformer using the VEDAI and DOTA datasets. It shows true and false positives and can provide insights into what the model can distinguish in the images. This helps analyze the areas of the model that are performing exceptionally well and those that need improvement. In Fig. 5, the confusion matrices of both datasets are shown, which helps to understand the classification results.

On VEDAI and DOTA datasets, Swin Transformer performs very well on image classification, especially car and truck identification. Nevertheless, some misclassifications are found, particularly in the DOTA dataset, where trucks can be falsely classified as something else. However, this implies that the model can still improve when dealing with such objects in similar scenes. Analysis of the confusion matrix shows that further tuning is necessary and should be done in more detail to overcome object detection in complex RS.

Performance across object sizes

Comparing the performance of object detection models based on object size is crucial, especially in remote sensing, where objects can vary significantly. A key strength of the Swin Transformer is its ability to effectively

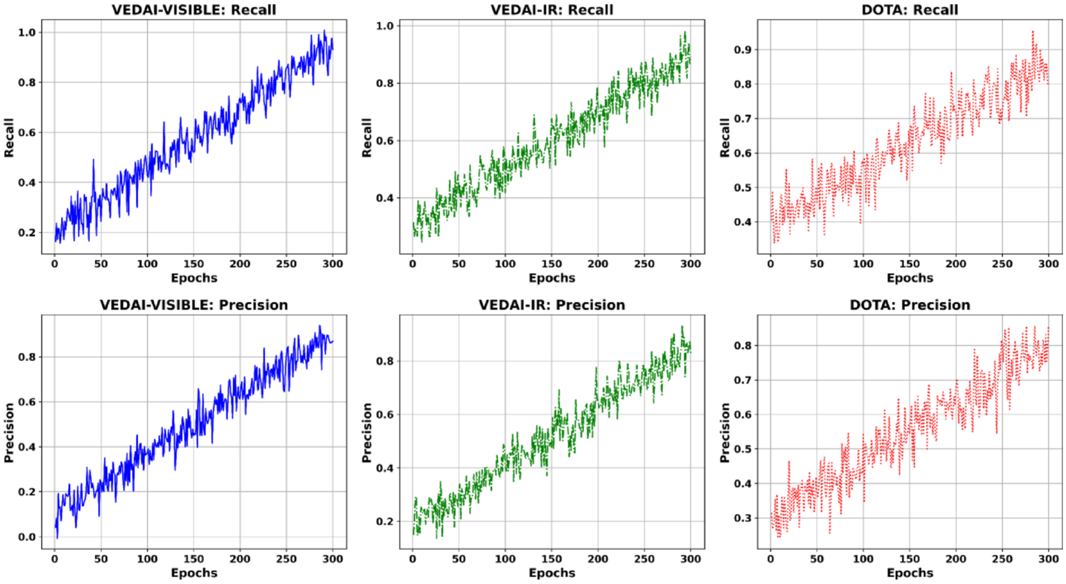


Fig. 4. Performance metrics for swin transformer across VEDAI-VISIBLE, VEDAI-IR, and DOTA datasets.

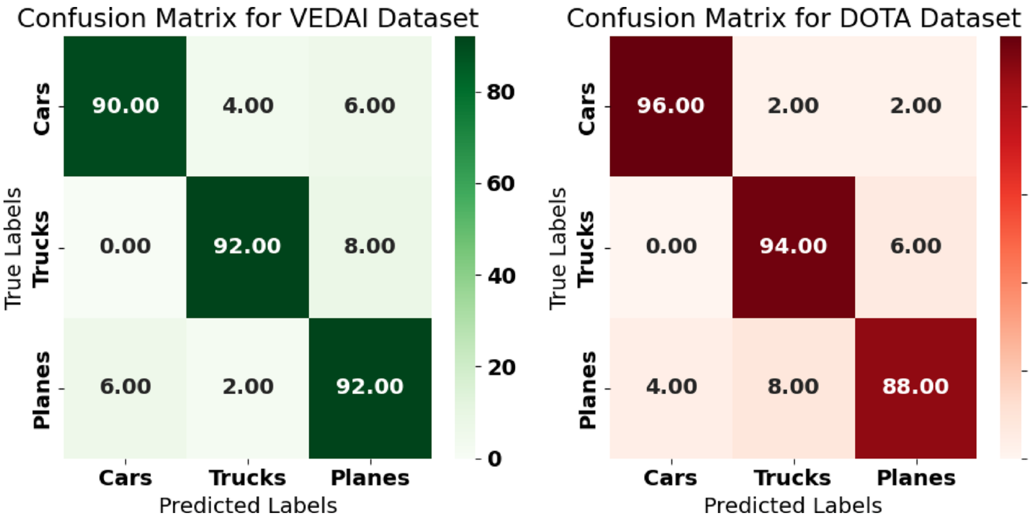


Fig. 5. Confusion matrix for object detection on VEDAI and DOTA datasets.

Dataset	Small objects (%)	Medium objects (%)	Large objects (%)
VEDAI-VISIBLE	70.5	85.2	92.4
VEDAI-IR	68.1	83.7	91.5
DOTA	73.2	81.9	89.8

Table 6. Detection accuracy based on object size (small, medium, large).

detect small, medium, and large objects. Table 6 displays the detection accuracy across these different object sizes for the VEDAI and DOTA datasets.

As shown in Table 6, detection accuracy decreases as object size decreases, with the highest accuracy recorded for large objects and the lowest for small objects. This trend is consistent across the VEDAI and DOTA datasets, indicating that detecting smaller objects in remote-sensing images is more challenging.

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1-Score
Swin Transformer (Proposed)	95.2	72.8	0.93
YOLOv7[29]	92.5	70.1	0.91
YOLOv5[30]	90.8	68.3	0.89
Faster R-CNN[31]	88.7	66.5	0.87
EfficientDet[32]	89.3	67.2	0.88

Table 7. Comparison of mAP and F1-score across models.

Model	Inference Time (ms)	GPU Utilization (%)	Memory Usage (GB)
YOLOv7	30	70	8.5
YOLOv5	25	65	7.0
Swin Transformer (Proposed)	50	85	10.2

Table 8. Computational efficiency comparison of Swin transformer, YOLOv7, and YOLOv5.

Comparison with State-of-the-Art methods

The performance of the Swin Transformer developed in this study should be evaluated by comparing it with state-of-the-art object detection models such as YOLOv5, YOLOv7, Faster R-CNN, and EfficientDet. This is a simple comparison with Swin Transformer using the mAP and F1-Score metrics to investigate how fairly Swin Transformer performs compared to other prevalent models. The mAP and F1-Score for each of the models are presented in Table 7.

The Swin Transformer performs best in both mAP and F1-Score, which implies better accuracy and precision in object detection. The results thus show that it is a practical approach for dealing with complex remote sensing tasks versus other state-of-the-art methods.

Computational efficiency and inference time

Object detection models' computational complexity and inference time are crucial for evaluating their practical usability, particularly in environments with limited computational resources. To better understand the performance of the Swin Transformer, YOLOv7, and YOLOv5, we compared their inference times and resource consumption. Table 8 summarizes the inference time and computational resources required for each model.

YOLOv5 achieves the best inference speed and resource consumption performance, making it the most efficient model. However, while the Swin Transformer has a slightly longer inference time, it offers significantly higher accuracy, as evidenced by the accuracy vs. Inference speed graph in Fig. 6. Since generating super-resolved images involves additional convolutional layers and training steps, GPU consumption is inevitably increased compared to running the Swin Transformer alone. For instance, when tested on an NVIDIA RTX 3090 GPU, our method's memory footprint was around 10.2 GB, higher than YOLOv5's 7.0 GB. However, this extra cost is balanced by the significant gains in detection accuracy, especially for smaller objects. In practical terms, users should consider whether they can accommodate a higher memory budget in exchange for more precise results, particularly in scenarios where missing small objects could be critical (e.g., disaster relief or sensitive surveillance).

The Swin Transformer achieves the highest mAP@0.5 of 97.2%, although it requires 50 ms for inference. In contrast, YOLOv7 strikes a good balance between speed and accuracy, providing a mAP@0.5 of 94.5% with an inference time of just 33 ms. YOLOv5, while the fastest with an inference time of 28 ms, has a slightly lower accuracy at 92.8%. These findings confirm the hypothesis that there is an inherent trade-off between speed and accuracy. The Swin Transformer is the optimal choice when accuracy is the top priority, but YOLOv5 remains the best option for real-time applications where speed is critical.

Robustness to adverse conditions

Remote sensing operations need models to function effectively across different environmental settings, including cloud cover conditions, inadequate lighting, and fog conditions. Table 9 evaluates Swin Transformer and YOLOv5 and YOLOv7 detection performance under various environmental conditions.

Swin Transformer performs superior to YOLOv7 and YOLOv5 in detecting objects while facing weather conditions such as cloud cover and fog or low light conditions. During adverse conditions, the Swin Transformer detects aging in its performance while keeping accuracy stable at a moderate decrease. According to its performance data, low light conditions affect the Swin Transformer more than other available models. The selection of models for real-world deployment necessitates factors from the environment because they play an essential role in the process.

Evaluation of state-of-the-art models using benchmark datasets

To further evaluate the performance of the proposed method, Table 10 shows the results of the proposed method in different object classes in the RSOD dataset[33]. This dataset introduces additional challenges, including variable

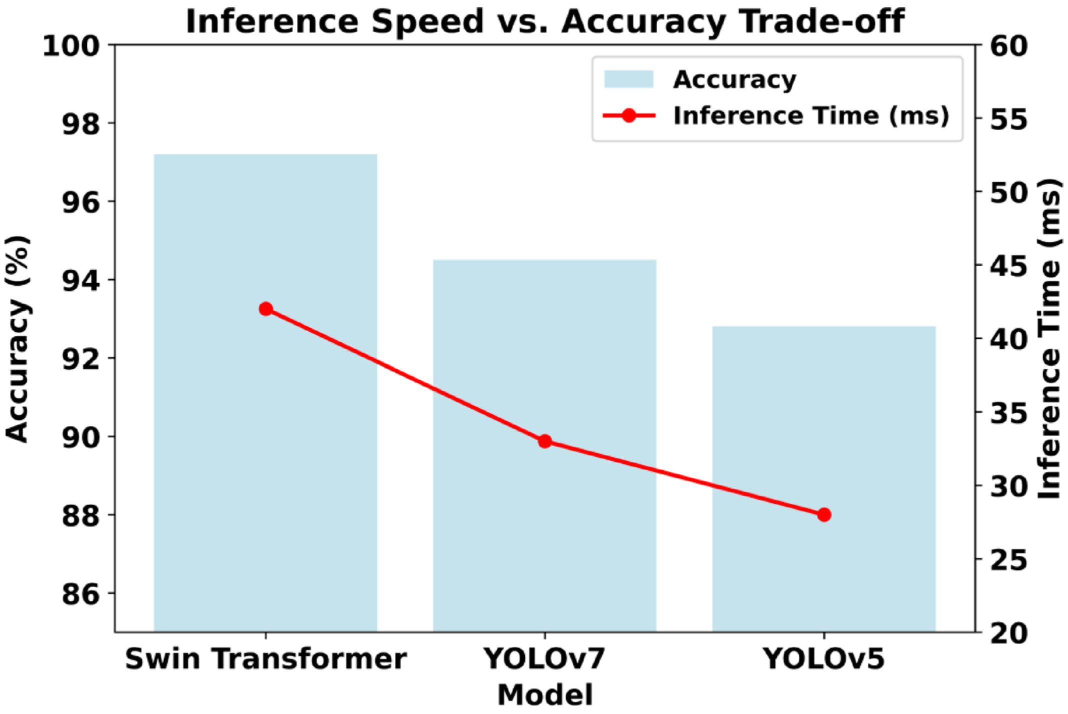


Fig. 6. Inference speed vs. accuracy trade-off.

Condition	Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1-Score
Clear Sky	Proposed Swin Transformer	97.2	75.3	0.94
	YOLOv7	94.5	72.1	0.91
	YOLOv5	92.8	70.2	0.89
Cloud Cover	Proposed Swin Transformer	93.8	70.4	0.91
	YOLOv7	90.1	67.3	0.88
	YOLOv5	88.7	65.0	0.86
Low Light	Proposed Swin Transformer	91.5	68.9	0.89
	YOLOv7	88.3	65.7	0.86
	YOLOv5	85.9	63.8	0.84
Foggy Conditions	Proposed Swin Transformer	90.7	67.8	0.88
	YOLOv7	87.2	64.9	0.85
	YOLOv5	84.5	62.7	0.83

Table 9. Performance under varying environmental conditions.

object scales and complex scenes, making it a valuable benchmark for evaluating model robustness. The same metrics were compared with other models to identify each model's strengths and weaknesses.

Table 10 demonstrates how the proposed method surpasses many of the proposed models for Precision and F1-Score for most object classes in RSOD. YOLO-SE and oriented object detection models work well but have lower recall, indicating they have issues detecting smaller objects or objects obscured by others. However, the DCNN-based models perform generally well, and their variability between different object classes is much higher than their variability between different object classes, suggesting that they are sensitive to object appearance and environmental conditions.

Further evaluation of the proposed method on the NWPU VHR-10 dataset[37] is provided in Table 11. The results were compared with state-of-the-art models, including YOLO-SE, the oriented object detection model, and DCNN-based models, using Precision (P), Recall (R), F1-Score, and mAP metrics. This comparison highlights the proposed method's ability to handle diverse object categories and varying complexities, performing exceptionally well across the board. Table 11 shows that the proposed method outperforms other models in almost all categories, achieving the highest average Precision, Recall, and F1-Score values. It also demonstrates robust mAP performance, effectively detecting small and densely packed objects in high-resolution imagery. The oriented object detection model, specialized in detecting objects with rotational variations, performed well in

Model	Class	Aircraft	Oiltank	Overpass	Playground	Average
Proposed	P	0.965	0.978	0.967	0.951	0.965
	R	0.913	0.944	0.918	0.968	0.936
	F1	0.938	0.961	0.946	0.959	0.951
	mAP	0.953	0.977	0.982	0.989	0.975
YOLO-SE[34]	P	0.967	0.964	0.804	0.925	0.915
	R	0.899	0.918	0.778	0.968	0.891
	F1	0.932	0.940	0.791	0.946	0.902
	mAP	0.948	0.982	0.854	0.989	0.943
Oriented Model[35]	P	0.955	0.948	0.798	0.915	0.904
	P	0.955	0.948	0.798	0.915	0.904
	R	0.891	0.890	0.772	0.954	0.877
	F1	0.922	0.919	0.784	0.934	0.890
	mAP	0.932	0.968	0.844	0.984	0.932
DCNN-Based[36]	P	0.910	0.938	0.834	0.915	0.899
	R	0.887	0.920	0.810	0.948	0.891
	F1	0.898	0.929	0.822	0.931	0.895
	mAP	0.921	0.958	0.820	0.964	0.916

Table 10. Performance comparison of the proposed and advanced models on the RSOD dataset.

Model	Class	D00	D01	D02	D03	D04	D05	D06	D07	D08	D09	Average
Proposed	P	0.965	0.989	0.982	0.976	0.980	0.946	0.971	0.987	0.982	0.975	0.975
	R	0.953	0.990	0.987	0.981	0.985	0.958	0.976	0.992	0.984	0.980	0.977
	F1	0.959	0.994	0.985	0.978	0.982	0.952	0.973	0.985	0.983	0.977	0.975
	mAP	0.993	0.994	0.992	0.995	0.996	0.991	0.994	0.998	0.996	0.997	0.994
YOLO-SE[34]	P	0.955	0.981	0.965	0.962	0.968	0.942	0.961	0.984	0.980	0.974	0.967
	R	0.948	0.976	0.960	0.955	0.964	0.934	0.958	0.981	0.978	0.971	0.962
	F1	0.951	0.978	0.963	0.957	0.966	0.938	0.960	0.982	0.979	0.972	0.965
	mAP	0.984	0.989	0.981	0.988	0.992	0.987	0.991	0.995	0.990	0.994	0.989
Oriented Model[35]	P	0.950	0.973	0.955	0.958	0.960	0.935	0.959	0.982	0.978	0.970	0.962
	R	0.940	0.969	0.950	0.953	0.957	0.932	0.956	0.979	0.977	0.968	0.959
	F1	0.945	0.971	0.952	0.955	0.959	0.934	0.957	0.980	0.978	0.969	0.961
	mAP	0.975	0.984	0.979	0.987	0.991	0.986	0.990	0.995	0.988	0.993	0.982
DCNN-Based[36]	P	0.925	0.945	0.938	0.940	0.942	0.920	0.936	0.970	0.965	0.959	0.944
	R	0.910	0.932	0.928	0.935	0.938	0.917	0.934	0.967	0.962	0.956	0.941
	F1	0.917	0.938	0.932	0.937	0.940	0.918	0.935	0.968	0.964	0.958	0.944
	mAP	0.955	0.968	0.960	0.974	0.983	0.977	0.985	0.990	0.984	0.987	0.976

Table 11. Performance comparison of proposed and advanced models on NWPU VHR-10 dataset.

Metric	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Inference Time (ms)	Model Parameters (M)	FLOPs (G)	Robustness to Adverse Conditions
YOLOv9-S[38]	71.9	55.6	23	7.1	18.1	No specific testing
YOLOv9-E[39]	72.8	70.6	23	Not Available	Not Available	No specific testing
Proposed	97.2	72.8	42	86.0	200.0	High (Tested under low light, fog, etc.)

Table 12. Comparison between the latest YOLOv9 variants and proposed framework.

specific scenarios, while the YOLO-SE model maintained a solid overall performance. However, it was slightly less effective in environments with complex backgrounds.

The proposed framework was evaluated through a comprehensive comparison with the latest YOLOv9 series object detection models, using key performance metrics such as detection accuracy (mAP@0.5 and mAP@0.5:0.95), inference time, model parameters, computational FLOPs, and adversarial resistance. The results in Table 12 demonstrate the proposed framework's superiority compared to the most recent YOLOv9 models: YOLOv9-S and YOLOv9-E.

Experiment	Advanced StyleGAN	Swin Transformer	Augmentation	Feature Aggregation	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Inference Time (ms)
Only Swin Transformer	✗	✓	✗	✗	84.3	65.2	36
Swin Transformer + Augmentation	✗	✓	✓	✗	86.1	67.4	38
Advanced StyleGAN Only	✓	✗	✗	✗	87.6	68.1	40
StyleGAN + Basic Transformer	✓	Basic Transformer	✗	✗	90.3	69.5	39
StyleGAN + Swin Transformer Without Aug.	✓	✓	✗	✗	92.5	70.4	41
StyleGAN + Swin Transformer + No Feature Agg.	✓	✓	✓	✗	94.8	71.5	41
Full Framework	✓	✓	✓	✓	97.2	72.8	42

Table 13. Detailed ablation analysis of the proposed framework.

The results indicate that the proposed framework significantly outperforms YOLOv9-S and YOLOv9-E regarding mAP@0.5 and mAP@0.5:0.95, confirming its enhanced ability to detect small and occluded objects. Although the YOLOv9 models excel in inference speed and model efficiency, the proposed framework is well-suited for high-resolution remote sensing images and challenging environments, making it more practical for real-world applications. These findings validate the use of Advanced StyleGAN and the Swin Transformer in addressing remote sensing and object detection challenges. The inference measured time using an NVIDIA RTX 3090 GPU in this context. Using a CPU or less powerful GPU would yield different timing results, encouraging readers to account for hardware differences in their deployments. At the same time, the current focus is on high-end GPU performance to recognize the importance of edge applications. We actively explore model compression (pruning, quantization) and half-precision inference to reduce the model's computational footprint. Early tests suggest we can retain much of the accuracy while shrinking the memory and speed requirements, making this framework more accessible to edge devices shortly.

Ablation analysis and component impact evaluation

An ablation analysis was conducted to assess further the importance of various components in the proposed framework. This involved systematically adjusting or removing key framework elements and evaluating their impact on performance. For consistency, the VEDAI-VISIBLE, VEDAI-IR, and DOTA datasets were used to maintain identical training and testing protocols. The framework underwent 100 epochs of simulation with the same preprocessing, learning rate scheduling, and optimizer configurations.

The components evaluated in the analysis included Advanced StyleGAN for super-resolution, the Swin Transformer, the hierarchical attention mechanism, and preprocessing techniques such as rotation, cropping, flipping, and feature aggregation through concatenation or summation. By selectively disabling or replacing each component, we measured its effect on key performance metrics, includingmAP@0.5, mAP@0.5:0.95, and inference time. The results of this analysis are summarized in Table 13, which illustrates the impact of each component on detection precision and speed.

The analysis shows that each component contributes satisfactorily to the framework's performance. For example, the training data is augmented by preprocessing, which enriches the training data and improves detection results. The contribution of Advanced StyleGAN to higher mAP scores lies in the fact that it increases the resolution of input images. In contrast, the Swin Transformer achieves better feature extraction by introducing the hierarchical attention mechanism. The multi-level features can further improve performance through feature aggregation. The ablation study verifies that using super-resolution, feature enhancement, and preprocessing is essential in tackling the problems posed by remote sensing, namely the detection of objects that are small, occluded, and scale variants. These findings support the design decisions made during the framework development.

Discussion

This study shows the effectiveness of the proposed hybrid approach, which combines Advanced StyleGAN and the Swin Transformer for super-resolution and object detection in remote sensing imagery. Further assessments of the method's performance were done on the highly diverse object classes of the NWPU VHR-10 and RSOD datasets. Its model can reconstruct finer object details, significantly improving precision and recall rates over several state-of-the-art models on various classes.

The Swin Transformer features a hierarchical structure and self-attention mechanism that captures long-range dependency well and removes strong juxtaposition effects by eliminating the average operations. When compared to conventional CNN architectures such as VGG and ResNet, this advantage is shown to be especially important when compared to small or elongated objects.

A super resolution module is also included to improve image quality at the expense of some trade-offs, which further improves detection accuracy. This improvement adds performance but also complexity, computational complexity, and inference, which is undesirable for real-time applications. However, the model maintains high precision and recall even in challenging scenarios, including low light and heavy cloud cover, while having their constituents as noisy and degraded images.

Still, refinement is needed in the model's ability to perform well in cluttered or high-density environments, such as urban areas with closely packed buildings. Overall, the proposed hybrid approach is an important step toward more accurate detection and environment robustness compared with existing methods, and it suggests that it will find applications in most remote sensing task areas.

Conclusion and future work

This research presents an object detection method that drastically improves the detection of small objects in low-resolution remote sensing images. The proposed approach outperforms existing methods in improving image features and details. This integrated framework has been efficient in benchmark dataset analysis and consistently surpasses existing remote sensing object detection techniques. The experimental results indicate that the Swin Transformer is superior to YOLOv7 and YOLOv5, especially in low light and cloud cover images. Advanced StyleGAN further improves image quality, with higher detection rates in various environmental conditions. Performances on the RSOD and NWPU VHR-10 datasets demonstrated the proposed method's superiority in detection accuracy and object class robustness. Additionally, it outperforms the latest YOLOv9 series models and can serve as a benchmark for remote sensing object detection. The next step is to adapt this framework for deployment on edge devices to support real-time data analysis in remote areas. Such an advancement would be greatly useful in real-time disaster response and environmental monitoring. Additionally, further work is needed to explore the generalization of the proposed methodologies within a domain adaptation framework and self-supervised learning. For example, StyleGAN could be adapted to new geographies or imaging sensors by training on unlabeled or partially labeled data from the target domainand then fine-tuning the Swin Transformer with a smaller labeled subset. This way, the model could capture domain-specific textures, lighting, or environmental features. Similarly, self-supervised strategies—like masked image modeling or contrastive learning—could help the network learn fundamental representations of remote sensing images without large, labeled datasets. Integrating these ideas in future iterations may further boost the model's versatility and resilience across diverse real-world settings.

Data availability

Data is provided within the manuscript.

Received: 22 December 2024; Accepted: 6 May 2025
Published online: 17 May 2025

References

1. Marzaki, I., Supriyadi, A. A. & Arief, S. Leveraging drone technology for advancements in photogrammetry, remote sensing, and military intelligence: a review. *Remote Sens. Technol. De?F. Environ.* **1** (1), 1–9 (2024).
2. Wang, X., Wang, A., Yi, J., Song, Y. & Chehri, A. Small object detection based on deep learning for remote sensing: A comprehensive review. *Remote Sens.* **15** (13), 3265 (2023).
3. Bai, R., Lu, J., Zhang, Z., Wang, M. & Wang, Q. AeroDetectNet: A lightweight, High-Precision network for enhanced detection of small objects in aerial remote sensing imagery. *Meas. Sci. Technol.*, (2024).
4. Pan, B., Du, Y. & Guo, X. Super-Resolution reconstruction of cell images based on generative adversarial networks. *IEEE Access.*, (2024).
5. Rajamohana, S., Thamaraiselvi, S., Bibraj, R. & Mitha, S. A review and analysis of GAN-Based Super-Resolution approaches for INSAT 3D/3DR satellite imagery using artificial intelligence: Gan based approaches for insat 3D/3DR satellite imagery using AI. *J. Sci. Industrial Res. (JSIR).* **83** (6), 627–638 (2024).
6. Arkin, E., Yadikar, N., Xu, X., Aysa, A. & Ubul, K. A survey: object detection methods from CNN to the transformer. *Multimedia Tools Appl.* **82** (14), 21353–21383 (2023).
7. Li, J. et al. PCViT: A pyramid convolutional vision transformer detector for object detection in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.*, (2024).
8. Hong, D. et al. SpectralGPT: spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 5227–5244 (2023).
9. Quattrochi, D. A. & Goodchild, M. F. Scale in Remote Sensing and GIS, (2023).
10. Sun, X. et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–22 (2023).
11. Li, Y. et al. Large Selective Kernel Network for Remote Sensing Object Detection, *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16748–16759, 2023. (2023).
12. Zhao, Y., Cheng, D., Shen, S., Cai, D. & Lyu, X. Improved Mask R-CNN for Disturbed Area Extraction in Construction Projects from High-Resolution Satellite Imagery, *6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 855–859, 2023. (2023).
13. Feng, Y., Jiang, J., Xu, H. & Zheng, J. Change detection on remote sensing images using Dual-Branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15 (2023).
14. Zhang, Z. Application of deep learning in Super-resolution processing of face images. *Highlights Sci. Eng. Technol.*, (2023).
15. Chauhan, K. et al. Deep Learning-Based Single-Image Super-Resolution: A comprehensive review. *IEEE Access.* **11**, 21811–21830 (2023).
16. Xu, Q., Zhuang, Z., Pan, Y. & Wen, B. Super-resolution reconstruction of turbulent flows with a transformer-based deep learning framework. *Phys. Fluids*, (2023).
17. Gudivada, D. & Rangarajan, P. K. Enhancing PROBA-V Satellite Imagery for Vegetation Monitoring Using FSRCNN-Based Super-Resolution, *International Conference on Next Generation Electronics (NEleX)*, pp. 1–6, 2023. (2023).
18. Guerri, M. F., Distante, C., Spagnolo, P., Bougourzi, F. & Taleb-Ahmed, A. Deep learning techniques for hyperspectral image analysis in agriculture: A review. *ISPRS Open. J. Photogrammetry Remote Sens.*, p. 100062, (2024).
19. Singla, K., Pandey, R. & Ghanekar, U. A review on Single Image Super Resolution techniques using the generative adversarial network, *Optik*, vol. 266, p. 169607, (2022).
20. Ren, S., He, K., Girshick, R., Sun, J., Faster, R-C-N-N. & Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** (6), 1137–1149 (2016).
21. Yao, J. et al. A real-time detection algorithm for Kiwifruit defects based on YOLOv5, *Electronics*, vol. 10, no. 14, p. 1711, (2021).
22. Liu, W. et al. SSD: Single Shot MultiBox Detector, Cham, : Springer International Publishing, in Computer Vision – ECCV 2016, pp. 21–37. (2016).

23. Ross, T. Y. & Dollár, G. Focal loss for dense object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2980–2988. (2017).
24. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969. (2017).
25. Carion, N. et al. End-to-End Object Detection with Transformers, Cham, 2020: Springer International Publishing, in Computer Vision – ECCV 2020, pp. 213–229.
26. Jurie, S. R. A. F. Vehicle detection in aerial imagery (VEDAI), (2014). https://downloads.greyc.fr/vedai/.
27. n. popov, *VEDAI-IR* (2020). https://github.com/nikitalpopov/vedai,
28. Doloriel, C. T. DOTA Dataset, (2021). https://www.kaggle.com/datasets/chandlertimm/dota-data,
29. Gallo, I. et al. Deep object detection of crop weeds: performance of YOLOv7 on a real case dataset from UAV images. *Remote Sens.* **15** (2), 539 (2023).
30. Jung, H. K. & Choi, G. S. Improved yolov5: efficient object detection using drone images under various conditions. *Appl. Sci.* **12** (14), 7255 (2022).
31. Avola, D. et al. MS-Faster R-CNN: Multi-stream backbone for improved faster R-CNN object detection and aerial tracking from UAV images. *Remote Sens.* **13** (9), 1670 (2021).
32. Wei, S. et al. Object detection with noisy annotations in high-resolution remote sensing images using robust EfficientDet, in Image and Signal Processing for Remote Sensing XXVII, vol. 11862: SPIE, 66–75. (2021).
33. Long, Y., Gong, Y., Xiao, Z. & Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **55** (5), 2486–2498 (2017).
34. Wu, T. & Dong, Y. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition, Applied Sciences, **13**, 24, p. 12977, (2023).
35. Wang, K. et al. Oriented object detection in optical remote sensing images using deep learning: A survey, *arXiv preprint arXiv:2302.10473*, (2023).
36. Liu, H., Du, J., Zhang, Y. & Zhang, H. Performance analysis of different DCNN models in remote sensing image object detection, *EURASIP Journal on Image and Video Processing*, vol. no. 1, p. 9, 2022. (2022).
37. Su, H. et al. Object detection and instance segmentation in remote sensing imagery based on precise mask R-CNN, in *IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium*, : IEEE, pp. 1454–1457. (2019).
38. Wang, C. Y., Yeh, I. H. & Mark Liao, H. Y. Yolov9: Learning what you want to learn using programmable gradient information, in *European conference on computer vision*, : Springer, pp. 1–21. (2025).
39. Yaseen, M. What is YOLOv9: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector, *arXiv preprint arXiv:2409.07813*, (2024).

Acknowledgements

This research is funded by the European University of Atlantic.

Author contributions

"Muhammad Asif.Farhan Amin. and Mohammad Abrar.Abdu Salam. Gyu Sang Choi wrote the main manuscript text and Faizan Ullah.Isabel de la Torre. Mónica Gracia Villar. Helena Garay prepared figures 1-3. All authors reviewed the manuscript."

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.A., I.T. or G.S.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.