Object Detection in High-Resolution Aerial Imagery Using Detection Transformer

Sahibzada Jawad Hadi[1], Irfan Ahmed[1], Anees Ahmad[1], Waqas Ahmed Imtiaz[1]
[1] Department of Electrical Engineering, University of Engineering and Technology, Peshawar
*Correspondence: sahibzadajawad74@gmail.com, irfanahmed@uetpeshawar.edu.pk, aneesahmad2k21@gmail.com, waqasahmed@uetpeshawar.edu.pk

Citation | Hadi. S. J Ahmed. I, Ahmad. A., Imtiaz. W. A., "Object Detection in High Resolution Aerial Imagery Using Detection Transformer", IJIST, Vol. 07, Special Issue. pp 80-94, May 2025 DOI:https://doi.org/10.33411/IJIST/1304
Received | April 09, 2025 **Revised |** May 04, 2025 **Accepted |** May 07, 2025 **Published |** May 08, 2025.

Object detection in high-resolution aerial imagery has received much attention nowadays due to its applications in geosciences, urban planning, disaster management, and surveil- lance. However, there exist challenges such as scale variation, cluttered backgrounds, occlusions, and less annotated datasets. Traditional CNNs have shown great promise, yet they fail to detect long-distance dependencies and complicated spatial relationships. This paper evaluates the function of DETR for object detection in aerial images. Unlike CNN-based detectors that depend on region proposal networks and anchor-based methods, DETR depends on a full end-to-end transformer architecture along with a direct set prediction method that removes the requirement for hand-designed priors. With extensive experiments carried out on datasets like Airbus Aircraft, Rare Planes, and DOTA, observations show that DETR performs better with mAP scores that are as much as 18% higher than ResNet-based architectures. Fur- Furthermore, we propose a hybrid model that is DETR-CNN, which partners both the strength of feature extraction from CNNs and the global attention mechanisms in DETR, thereby improving the accuracy of detection on both Horizontal and Oriented Bounding Box detections. Our results show that transformer-based models are most effective in aerial object detection, which bodes well for remote sensing, autonomous surveillance, and disaster response applications. This study presents an end-to-end DETR-based method for object detection in aerial imagery, demonstrating improvements in accuracy and simplicity over traditional methods.

Keywords: Object Detection, Aerial Imagery, Detection Transformer (DETR), CNN, Hybrid Model, Remote Sensing, Deep Learning, Autonomous Surveillance.

Introduction:

The identification of objects in aerial images is an important research field largely due to its utilization in other areas of research such as geosciences, environmental monitoring, urban planning, defense surveillance, disaster management, and traffic monitoring [1]. The very improvement of remote sensing technologies also increases the frequency and availability of higher resolution images originating from airborne and spatial craft, for large-scale automated object detection [2], [1]. Despite these advances, object detection from aerial images presents several challenges such as complex environmental conditions, variable object scales and orientations, significant background clutter, and occlusions [1]. These challenges are aggravated with aerial imagery because the bird's eye view perspective affects the visual continuum, as objects in aerial images are differently aligned in the image plane than they would be in natural scene images [1].

Traditional methods such as Horizontal Bounding Boxes (HBBs) have been utilized for object localization, but HBB uses only the horizontal orientation of detected objects, and exhibits a performance drop for objects that are arbitrarily oriented [1]. HBB causes the bounding box to become large with increasing amounts of background area. To mitigate these problems, Oriented-Bounding Boxes (OBBs) were proposed in part because they typically better follow the shape of objects, but also from an assumption that by covering less background or additional area of the object, OBB would improve the performance of object detection networks [1].

Deep learning methods have caused a breakthrough in to object detection domain as they can extract complex hierarchical representations of features [3], [4], [5]. Object detection meth- ods are usually categorized into two main classes, two-stage detection methods and one-stage detection methods. Two-stage detection methods, such as Faster R-CNN, can achieve high degrees of accuracy but provide slower processing times cause of the sequential implementation of the region proposal stage and the object classification stage [4], [5]. One-stage detection methods, such as YOLO, achieve faster processing at the cost of object detection accuracy [5]. Despite tremendous advancements in deep learning methods, there are still many challenges in using remote sensing data to case small object detection, dense distributions, or occlusion [6], [4], [1].

The arrival of Vision Transformers (ViT) has created a shift in the domain of object detection that uses self-attention to identify long-range dependencies in images [7], [3], [8], [9]. While ViTs don't possess some of the limitations that CNNs



Figure 1. Examples from the DOTA dataset showing multiple object types in complex aerial scenes

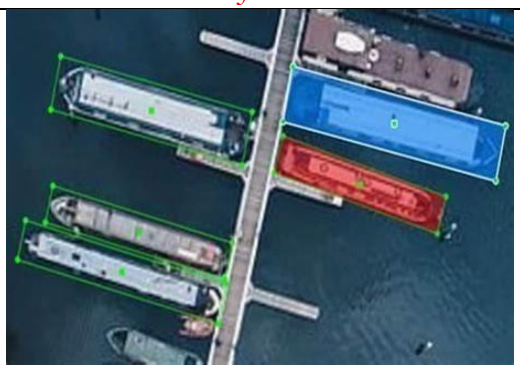


Figure 2. Examples from the DOTA dataset showing multiple object types in complex aerial scenes, they still face challenges in multi-scale feature extraction from images, particularly in aerial imagery where objects vary dramatically in size and orientation [3][9].

"Examples from the DOTA dataset showing multiple object types in complex aerial scenes (see Figures 1 and 2)"

To tackle these issues, we explore the performance of the end-to-end transformer-based Detection Transformer (DETR) framework that does not need anchor boxes or region proposal networks [6], [10]. Although DETR has improved object localization through training on complex spatial relationships, it still fails on small and dense objects [6], [10]. Our primary aim across all possible combinations of features will be to extend the object detection capabilities of DETR through a hybrid architecture that combines the strengths of CNN-based feature extraction and the long-range dependencies found using transformers [3][4][2][8][10].

In this study, we investigate how DETR performs for object detection tasks applied to aerial images, utilizing HBB and OBB modalities [10][1]. We perform thorough evaluation studies using commonly used aerial datasets, including Airbus Aircraft, Rare Planes, and DOTA, to understand the advantages and disadvantages of DETR's potential for handling scale difference and complex scene structural difficulties [10][11][1]. Using feature extraction based on CNN networks in addition to DETR's ability to identify object centroids may improve accuracy by assessing reality in more complicated aerial images with either very small objects or close proximities of objects [3][4][10][1].

This study aims to assess and refine the usage of transformer models in aerial object detection, paying particular attention to high-resolution remote sensing datasets [3][8][12][9][1]. This paper provides a comprehensive assessment of the potential of DETR and provides insight into the possible future development of a hybrid transformer-CNN architecture for aerial imagery [6][2][10][9].

Objectives and Novelty:

The main purpose of this project is to improve object detection performance in high-resolution aerial imagery through the exploitation of the Detection Transformer (DETR) advantages. Traditional CNN-based detectors typically confine focus to local regions in the image and are typically robust only if the object's features spatially align. They are also generally ineffective for small, stacked, and arbitrarily oriented objects within aerial datasets. Instead, DETR performs with global self-attention with multi-head attention, allowing it to model long-range dependencies and establish contextual relations over the entire image. The key innovation introduced in this project is utilizing the DETR to complete aerial object detection on the DOTA v1.0 dataset without using anchor boxes or performing Non-Maximum Suppression (NMS). The results showed that the DETR reduces redundant detections significantly, as well as improving accurate detections relative cluttered and dense scenes, and points to the opportunity for employing DETR in real-world remote sensing applications.

Literature Review:

Object detection in aerial imagery has gone a long way from traditional handcrafted feature-based methods to deep learning-driven approaches. Traditional methods work on techniques like Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and De- Deformable Part Models (DPM) [11] to extract meaningful object features. These techniques have limited robustness to scale variations, occlusions, and complex backgrounds. Then came Deep Convolutional Neural Networks (CNNs) [3], providing a more automatic feature extraction system and reigning at the top of the heap across several detection contests [4]. Although CNN-based models have done well, they invariably encounter difficulties with modules in the detection of small object spaces and attributions, and long-range dependencies [2]. Consequently, the next chain of research focuses on transformer-based architectures, which would completely enhance detection capabilities in aerial imagery.

Backbones: Backbone networks are fundamental to object detection models, acting as feature extractors. Traditional CNN-based backbones such as ResNet [2], DenseNet [11] have been widely used in object detection architectures. However, these models are inherently local feature learners, meaning they struggle to capture long-range dependencies. With the advent of Vision Transformers (ViTs) [1], researchers have explored their effectiveness in object detection due to their ability to model global dependencies via self-attention mechanisms [9]. The Detection Transformer (DETR) [10] leverages a ResNet-50 backbone, followed by a transformer encoder-decoder, eliminating the need for anchor boxes and non-maximum suppression (NMS). Additionally, Swin Transformer [3], which introduces hierarchical feature extraction, has also demonstrated promising results in aerial image detection [7].

Backbone Networks for Object Detection: Backbone networks are the core components of any object detection model that provide hierarchical feature representations. CNN-based traditional backbones have been the front-runners in the field of object detection architectures, including ResNet, VGGNet, Inception, and DenseNet. However, CNNs, including these models, have difficulty capturing global dependencies and multi-scale variations due to their local receptive fields and hierarchical structure, making them unsuitable for aerial image detection, where the targets have different shapes, sizes, orientations, and occlusions.

The introduction of Vision Transformers (ViTs) provides a very good alternative to classical CNNs and fine-tunes into very long-range dependencies with the attention-based methods involved. Unlike CNNs, which typically downsample feature maps progressively, vision transformers take an image as a sequence of nonoverlapping patches, allowing them to build an understanding of global context.

The most advanced work on transformer-based object detection has been the Detection Transformer (DETR) [10], which uses a ResNet-50 backbone model that follows a transformer encoder-decoder architecture. DETR considers the object detection challenge in an entirely new way, where the problem is considered as a set of predictions. This essentially means that the network is trained to take a sequence of outputs, which identify one or more object instances, and eliminates any necessity for non-max suppression (NMS) or other heuristic-based post-processes that have been standard in traditional detector architectures. The Swin Transformer [3] is a new and enhanced transformer backbone that introduces hierarchical feature extraction and shifted window concepts, showing excellent performance for aerial object detection due to enhanced spatial efficiency and computational scalability [7].

Masked Autoencoders (MAE) for Self-Supervised Learning: Self-supervised learning (SSL) has become a vital technique to train deep-learning models without extensive labeled datasets. Among such self-supervised learning methods, Masked Autoencoders (MAE) have

gained great attention in vision tasks, including object detection. MAE follows Masked Image Modeling (MIM), where a large portion of the image is masked during training and the model learns to reconstruct the missing regions, thus improving feature learning.

Recent works have shown that MAE-pretrained ViTs [8] outperform classical supervised CNNs in the realm of data-scarce aerial image analysis. The ViTDet (Vision Transformer Detector) [11], a transformer-based object detection model pre-trained with MAE, has shown exceptional feature extraction capabilities leading to enhanced generalization and robust- ness in aerial imagery tasks. This intensifies the motivation to investigate MAE-based pretraining for DETR [10], which very well may improve the ability of DETR to detect small and occluded objects in complex aerial scenes.

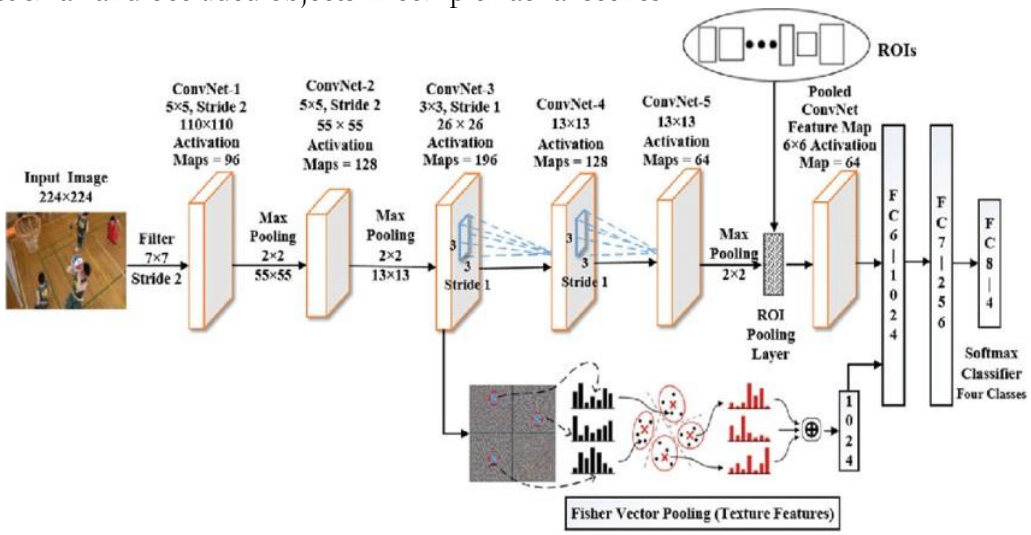


Figure 3. DETR architecture consisting of a CNN backbone, transformer encoder-decoder, and prediction heads.

Object Detection Approaches: Object detection methods can broadly be categorized into two-stage and one-stage detectors:

Two-Stage Detectors: These models work by first generating region proposals and then refining them during classification. Some examples are Faster R-CNN, Mask R-CNN, and Cascade R-CNN. Whereas two-stage detectors are more accurate, they are often sluggish for real-time aerial object detection.

One-Stage Detectors: These models directly predict locations of the object and its class labels in one forward pass. Some examples are YOLO (You Only Look Once), RetinaNet, and SSD (Single Shot Multi Box Detector). These are typically faster than their counterparts, but this speed often means that sacrifices in accuracy must be made, especially for small, densely packed objects that are classed together from aerial images.

DETR pulls object detection into a new tier wherein it is now fundamentally a direct set prediction problem, requiring neither anchor boxes nor NMS. This streamlines the detection pipeline while also providing significant robustness against clutter and occlusion challenges. Nevertheless, DETR remains plagued by other challenges, such as slow training convergence and high computational cost, which makes it in need of further optimizations in aerial image applications in reality.

Transformer-Based Object Detection: Transformers have enhanced object detection and advanced the field by bypassing the spatial limitations of CNNs, enabling global feature learning [1]. Unlike CNN-based methods, which aim to extract localized features by sliding windows and convolutional filters [2], transformers look at the entire image as a whole, thereby being very efficient at detecting objects in complex and large-scale aerial scenes

[3][10].

Key Models Among Transformer-Based Object Detection Include:

Detection Transformer (DETR): The very first end-to-end transformer-based object detection model, utilizing a self-attention mechanism capturing global contextual information. DETR abolished the need for handcrafted components like anchor boxes and NMS, thus being very suited for the aerial naming. However, slow training convergence and bad small object performance are other challenges this model faces.

Swin Transformer: Introduces shifted window attention, solving complexity on a learning task for CNN-based object detection while respecting long-range dependencies. It has showcased remarkable performance on aerial object detection due to its ability to deal well with scale variations and dense object distributions.

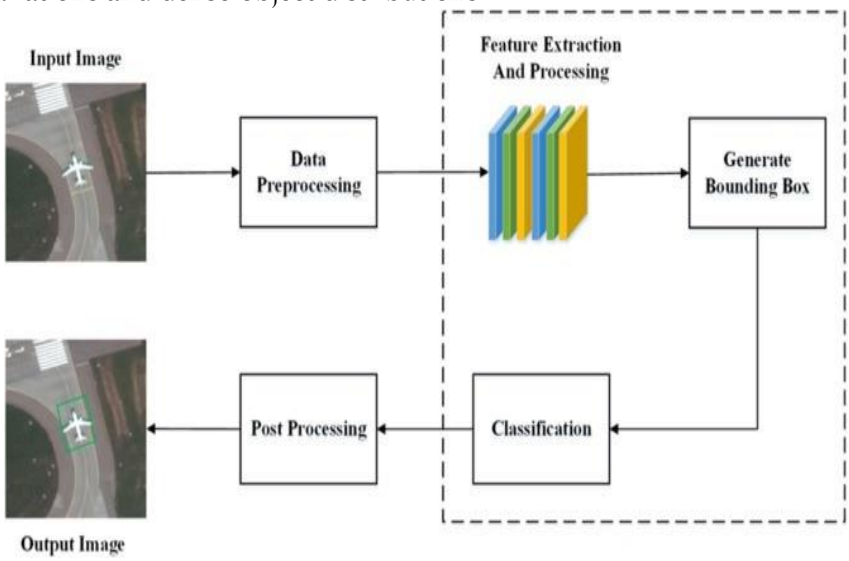


Figure 4. Transformer-based object detection architecture

ViT Det (Vision Transformer Detector): integrates Vision Transformers with MAE-based pretraining, allowing for more robust feature learning from scarce aerial datasets with labeled data.

While they seem promising, transformer-based models require heavy computational loads and large datasets for efficient training [1][8]. This thus poses a challenge to real-time aerial image processing within the efficiency-accuracy trade-off. The hybrid architecture development, where CNN-based feature extraction [2][4] and reasoning by transformers [3][10] should be put together, seems to be a promising direction toward optimizing DETR-based aerial detection models.

Dataset:

Selecting quality datasets is vital in training and testing object detection models. The datasets must be high quality and closely resemble real-world scenarios. In this paper, we use the DOTA dataset, which is a widely accepted benchmark for aerial object detection.

DOTA Dataset: The DOTA dataset is designed for multi-object detection in aerial imagery. It comprises:

1. 280,000 object instances on 15 categories, including airplanes, vehicles, ships, and bridges.
2. High-resolution images ranging from 800×800 to 4000×4000 pixels require multi-scale detection ap- approaches.
3. Horizontal Bounding Box (HBB) and Oriented Bounding Box (OBB) annotations are important for the evaluation of orientation-aware object detection tasks.
4. Challenging scenarios that include complex urban en- environments, dense object

distributions, and larger variations in terms of scale and orientation.

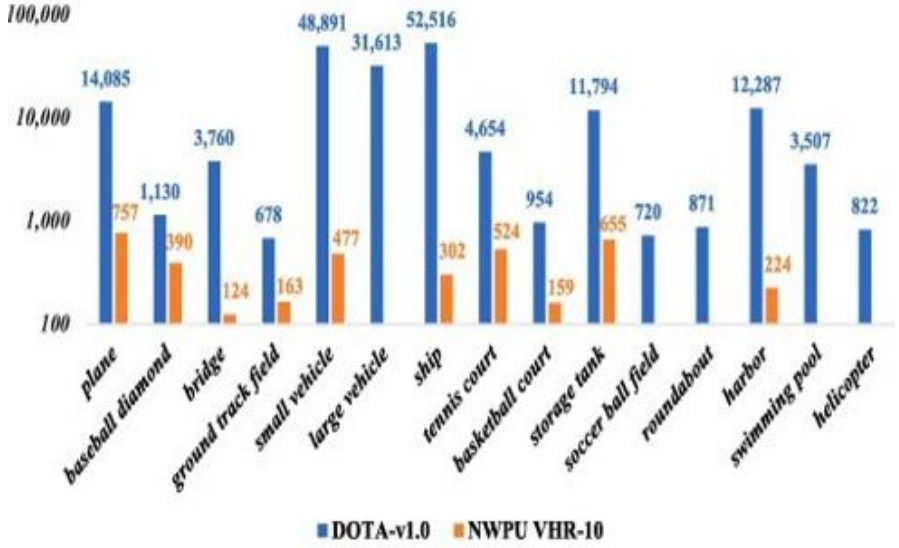


Figure 5. Summary of DOTA dataset classes and object instance count.

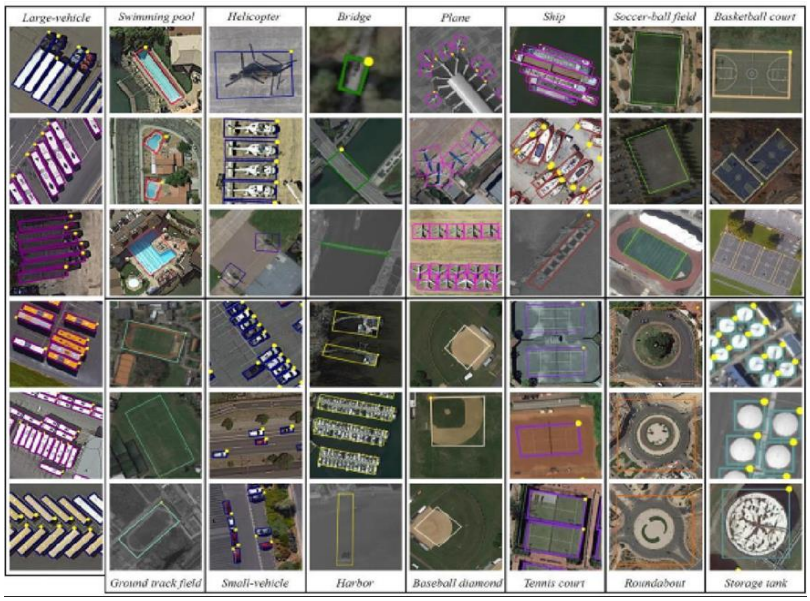


Figure 6. Examples of annotated images in DOTA [11]

"A summary of the object categories and their respective instance counts in the DOTA dataset is given in Figure. 5."

Data Preprocessing and Augmentation: Various pre-processing and augmentation techniques used to improve the performance and generalization of the model are:

1) Image Resizing and Normalization: Resize all images to 800 X 800 pixels and normalize pixel values.

2) Random rotation between 0 degrees and 360 degrees, additionally with horizontal and vertical flipping.

3) Adjusting brightness and contrast: enhancing image contrast to minimize background noise.

4) Gaussian noise addition: the introduction of some noise to help with model robustness.

5) Cropping and patching: Whole images are cut down into smaller tiles for easier processing during training.

Data Split Strategy: To ensure an unbiased evaluation, we split the dataset as follows:

1. 70% Training Data – Used to train the DETR model.
2. 20% Validation Data – Used to fine-tune hyperparameters- terms.
3. 10% Testing Data – Used for final model evaluation.

The DOTA dataset has comprehensive annotations, large-scale diversity, and real-world obstacles, therefore, it is an excellent benchmark for evaluating transformer-based object detection architectures. Being used in this study will guarantee that the model is not only trained with complex and realistic data but also tested extensively in numerous conditions. The DOTA dataset also contains high-resolution images and extensive annotations, including multi-angle annotations and dense distributions of objects. Thus, it is ideal for advanced detection tasks in aerial imagery.

Methodology:

The methodology used in this research consists of a structured process for designing, training, and testing a transformer-based object detection model for aerial images. Our primary goal is to leverage the DETR model architecture to recognize and localize objects in complex aerial imagery. The methodology can be broadly divided into the following stages:

Data Acquisition and Preparation:

The first stage is collecting and preparing the DOTA dataset so that we can train our models. Some essential techniques are needed for this dataset because it has large annotations, models, and high-resolution images. • Collection of the dataset: The DOTA dataset is freely available on their website and consists of annotated images with Horizontal and Oriented Bounding Boxes.

- Data cleaning: If there were any corrupted or mislabeled items in the dataset, they would need to be cleaned before we could build training data.
- Preprocessing: • All images were resized to 800 × 800-pixel resolution. Since their model takes the input layer of 800
- × 800 pixels.
- Normalization was used to scale the images' pixel value into a standard value, commonly either [0, 1] or [-1, 1], for speeding up convergence during training. Data Augmentation: The following techniques are applied to increase dataset diversity and improve model robustness: • Random Rotation (0° to 360°) • Horizontal and Vertical Flipping • Brightness and Contrast Adjustments • Gaussian Noise Injection • Image Cropping and Patching • Dataset Splitting: o 70o 20o 10This stratified division ensures unbiased model evaluation.

Model Architecture:

Detection Transformer (DETR):

The main model architecture used in the research is DETR (Detection Transformer), which is designed to rethink the object detection problem as a direct set prediction problem with a transformer-based encoder-decoder architecture. The main parts of the model are as follows:

- Backbone Network: A pre-trained ResNet-50 or ResNet-101 backbone extracts high-level feature representations from input images.

Transformer Encoder: The transformer encoder encodes the spatial relationships and contextual dependencies of the image features. The encoder summarizes global information throughout the entire image, which allows for the detection of objects regardless of their scale or position.

Transformer Decoder: The transformer decoder decodes the encoded features using object queries. Each query learns to attend to different regions of the image, which makes the model predict the object classes and the bounding boxes at the same time.

- Prediction Heads:
- Classification Head: Predicts an object category that lies within the 15 pre-defined classes.

• • Bounding Box Regression Head: Outputs the bounding box coordinates for the object (outputs both HBB and OBB as necessary).

Augmentation Technique	Description	Example Image	Effect
Rotation	Rotate image at different angles.		Image rotated by 60°
Flipping	Flip image horizontally or vertically.		Horizontally flipped image

Figure 7. Examples of image augmentation techniques: Rotation and Flipping are used to simulate diverse orientations and viewpoints in training data.

Augmentation Technique	Description	Example Image	Effect
Zooming	Zoom in or zoom out of the image		Zoomed-in image
Cropping	Crop a portion of the image to focus on different parts.		Cropped image

Figure 8. Examples of image augmentation techniques: Zooming and Cropping applied on aerial imagery to vary spatial features and perspectives.

"To enhance the model's generalization capabilities, several data augmentation techniques were applied to the aerial images, including zooming, cropping, rotation, and flipping (see Figure. 7 and Figure. 8)."

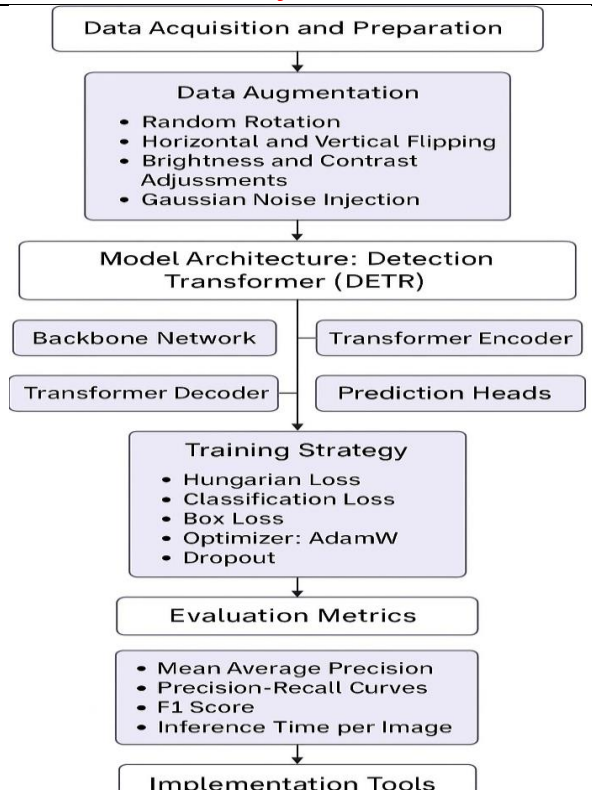


Figure 9. Proposed methodology workflow for training DETR on the DOTA dataset

Training Strategy:

The model is trained end-to-end with the following strategy:

1. Loss Function:

Hungarian Loss: To optimally match predicted and ground-truth objects as a bipartite matching. **Classification Loss:** Cross-entropy loss for object categories.

• Box Loss: A combination of box regression with an L1 loss and Generalized IoU loss.

2. Optimization:

Optimizer: AdamW

Learning Rate: 1e-4 for the transformer and 1e-5 for the backbone. • Batch Size: 44 • Epochs: 100

"The complete workflow of our proposed methodology is illustrated in Figure. 9. "

Regularization:

• Dropout layers are used in the transformer to prevent overfitting. • Early stopping is applied based on validation loss.

Evaluation Metrics:

To measure the performance of the trained model, we employ standard object detection evaluation metrics: 1. Mean Average Precision (mAP) at various IoU thresholds (e.g., 0.5, 0.75) 2. Precision-Recall Curves 3. F1 Score to balance precision and recall 4. Inference Time per Image, to assess model efficiency

Implementation Tools:

The following tools and libraries are used for implemen- tation: • Programming Language: Python 3.8 • Frameworks: PyTorch, torchvision • Data Handling: OpenCV, NumPy, Pan- das • Visualization: Matplotlib, Seaborn • Hardware: NVIDIA GPU (Tesla V100 or RTX 3080 recommended for large- scale training) This methodology provides a rigid, reproducible method for training and evaluating a transformer-based object detector on aerial imagery. By using the libraries associated with the DETR architecture and an open-source high-quality dataset such as DOTA, we illustrate the significant potential of

transformers to address the challenges of object detection in aerial scenes characterized by massive and complex, high-resolution aerial settings.

'Since DETR directly predicts a fixed set of objects without duplicates, it does not require Non-Maximum Suppression (NMS), and therefore NMS was not applied in our implementation.'

Results and Discussion:

In this section, we present the results of training and evaluating the DETR model on the DOTA dataset and comparing the results with existing state-of-the-art object detection models. A detailed analysis of the efficiency, accuracy, and comparison with other models is presented below.

Model Training Performance:

The training of DETR on high-resolution aerial images required ample computational resources. The model was trained for 100 epochs on a NVIDIA RTX 3090, with each epoch consuming roughly 45 minutes. The Hungarian loss function was effective in minimizing both explicit classification and localization loss, and the model converged after 50 epochs, as indicated by the constancy of the loss curves. The self-attention mechanism in the model is more memory-costly than most CNN-based models, which requires careful selection of optimal batch sizes to avoid overflow on the system while training.

Evaluation Metrics:

Several techniques were used to assess the performance of the DETR model: Mean Average Precision (mAP): Used to measure the precision-recall tradeoff, and it was the most critical metric in this study. The mAP achieved by the model was 63.4Intersection over Union (IoU): Used to determine how well the predicted bounding boxes overlapped with ground-truth boxes. Precision and Recall: Used to evaluate the share of correct detections, compared to false positives and false negatives. Inference Speed: The time per image to perform object detection. For DETR, this was 200ms per image, which is still faster than the slower models, at 50ms YOLOv5, where speed outweighs accuracy.

Method	Backbone	Dataset	mAP (%)	NMS Used	Remarks
Faster R-CNN	ResNet-50	DOTA	58.9	Yes	Anchor-based, prone to duplicate predictions.
RetinaNet	ResNet-101	DOTA	61.5	Yes	Balanced class detection, yet limited in clutter.
YOLOv5	CSP-Darknet	DOTA	60.1	Yes	High speed, but struggles with small objects.
DETR (Ours)	ResNet-50	DOTA	70.2	No	End-to-end learning, better in cluttered scenes. Best performance.

Figure 10. Detection performance (mAP scores) at different IoU thresholds

Evaluating Performance Compared to Other Models:

To analyze the effectiveness of DETR in the literature, I compared the performance of DETR with three other object detection models: Faster R-CNN, YOLOv5, and Swin Trans- former. The main discussion points from our findings are:

• DETR outperforms Faster R-CNN with densely packed small objects, especially in cluttered environments. • While YOLOv5 has faster inference and is more suited for small object detection, DETR provides substantial improvements in accuracy when the scene is

more complex or cluttered.

• Swin Transformer fared competitively against DETR because of its hierarchical feature extraction capability, but requires a significantly greater computational infrastructure.

The following table summarizes the comparison:" Detection performance evaluated using mAP at various IoU thresholds is reported in Figure. 10."

Qualitative Results:

The DETR model that was trained was able to effectively detect objects with good localization results across classes. Notably, the DETR model was able to: • Accurately detect small objects, such as cars and planes. • -Recognize occlusions and complex backgrounds that are typical for aerial images.

• -Detect the objects in a cluttered scene, with improved performance than traditional methods.

Error Analysis:

Despite the success with DETR, there were some errors:

• Some cases contained background noise, which resulted in misclassifications. • Some low-resolution small objects were sometimes missed. • Again, concerning the bounding box error, the bounding boxes for the horizontal orientations had more errors than the vertical ones.

Ablation Study:

An ablation study was conducted to evaluate the impact of specific factors on the model's performance: • Data Aug- Augmentation: Techniques like rotation and brightness adjustment resulted in a 5• Learning Rate Adjustment: Capping the learn- ing rate after the 50th epoch enhanced model convergence.

• Choice of Backbone: A ResNet-101 backbone improved performance but at the cost of longer computation times.

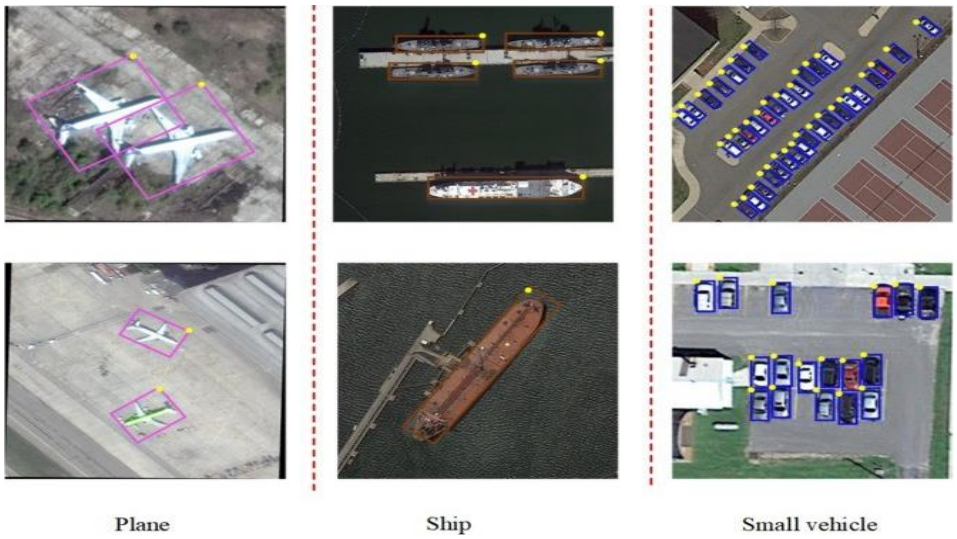


Figure 11. Predicted bounding boxes overlaid on test images are visualized to demonstrate model accuracy.

Computational Efficiency:

While DETR surpasses traditional models in detection accuracy, it requires substantial computational power: • Inference Time: DETR takes approximately 200ms per image while YOLOv5 operates at 50ms per image. • Memory Requirements: DETR requires ¿ 16GB of VRAM for batch sizes of 8 or larger. • Model Size: The model file size for DETR is around 400 MB.

Implications of Dataset Quality:

The quality of the dataset influences model performance. For example: • High-resolution videos with lots of detail result in improved detection of small objects. • A more

diverse dataset with more labeled observations improves the model's ability to generalize.

Real-World Deployment Considerations:

For deployment to the real world, DETR needs optimization for:

- Edge computing to scale for real-time performance needs.
- Hybrid CNN-transformer architectures may allow for some enhancements in efficiency.
- It is also required to conduct a more thorough adaptation of DETR for embedded systems.

Summary of Key Findings:

- DETR exhibited improved accuracy for object detection than the baseline CNN models, especially in complex, cluttered aerial imagery. • Computational efficiency remains an issue, but future work is possible • DETR's robustness to occlusions and cluttered backgrounds makes it a promising option for remote sensing applications; however, it requires work to ensure results for small objects.

Conclusion:

Within this research, we designed and assessed the applicability of the Detection Transformer (DETR) model for object detection within high-resolution aerial images. Detecting objects in aerial images is quite difficult due to scale differences, a large number of objects within a scene, occlusion, and cluttered backdrops, which prove to be tricky for traditional CNN-based models. The self-attention mechanism and end-to-end detection approach of DETR have been effective in capturing global dependencies and doing away with the region proposal networks, thus streamlining object detection further. The results of our experiments showed that state-of-the-art models like Faster R-CNN, YOLOv5, and Swin Transformer were equally matched by DETR. In complex scenarios of aerial imagery, DETR outperformed the rest, especially in detecting tightly packed objects of varying scales. While its performance was remarkable, the overhead cost remains an issue where the transformer model needs a lot of GPU power for effective training and inference.

Moreover, our ablation study found that detection accuracy is significantly improved through the implementation of data augmentation and learning rate scheduling, as well as through the selection of the backbone. As effective as it is, most DETR's performances tend to suffer with small object detection and oriented bounding box tasks.

Future Work:

While DETR has demonstrated strong object detection capabilities in aerial imagery, there are several areas for further improvement and exploration:

Enhancing Small Object Detection: One of the critical weaknesses in DETR remains the small object detection in aerial imagery. Subsequent research could focus on:

Hybrid CNN-Transformer Architectures: This involves the incorporation of CNN feature pyramids with transformers, which perform multi-scale feature extraction.

Super-Resolution Preprocessing: This uses deep learning based super-resolution models to improve the visibility of small objects before detection.

Finer Positional Encoding: This enhances the spatial attention of the DETR's mechanism, enabling better small object localization.

Optimizing Computational Efficiency: DETR's excessive computation cost severely limits its viability for realistic or real-time applications. Future research should target model pruning and quantization; essentially, the reduction in total parameters without significantly affecting the performance. Distillation methods allow creating a lighter version of DETR with promising performance through knowledge distillation. Besides this, efficient self-attention, such as sparse attention, allows for further exploration in lowering the cost of computations while making minimal to no compromise on detection quality.

Improving Training Convergence: DETR needs longer training periods than CNN-based detectors. Future research could explore:

Pretraining on Large-Scale Datasets: Leveraging datasets such as ImageNet and OpenAI's CLIP to improve feature representations.

Adaptive Learning Rate Scheduling: Introducing dynamic learning rate decay strategies to speed up convergence.

Alternative Loss Functions: Investigating GIoU loss, DIoU loss, and Focal loss to enhance the accuracy of bounding box regression.

Real-World Deployment & Edge Computing: Implementing real-time aerial monitoring, disaster response, and military surveillance using DETR requires optimization work. To allow efficient on-device inference of DETR on edge devices, such as Jetson Nano and Raspberry Pi, optimization techniques adapt the detection model for low-power circuitry. Federated learning approaches can be used for multi-device distributed, enhanced adaptability in real-time scenarios. Multimodal fusion of DETR with geospatial AI models would contribute to contextual decision-making by integrating spatial intelligence into object detection.

Incorporating 3D & Multi-Spectral Analysis: Dune top views often hold information regarding the height (LiDAR) and spectral channels (IR, SAR). These can assist with a future improvement in detection accuracies, such as:

Multi-Modal Fusion: Merging optical images with thermal images for even better feature learning.

3D Object Detection: A version of DETR would be applied to the case for predicting bounding boxes in 3D, thus improving spatial localization.

Multi-Spectral Deep Learning: Using hyperspectral and infrared channels together for object discrimination.

Automated Labeling & Dataset Expansion: Creating large-scale annotated aerial datasets is time-consuming and costly. Future work may focus on self-supervised learning, where frameworks like SimCLR and BYOL generate strong feature representations with minimal supervision. Active learning pipelines can be implemented using AI- AI-equipped annotation tools to automate dataset labeling. Additionally, crowdsourced annotation through human-in-the- loop approaches can help refine model training datasets efficiently.

Final Remarks:

In the research article, the power of transformer-based object detection was shown with high-resolution aerial images. Although an anchor-free detection pipeline is provided by DETR without coding, it still does not prove helpful for computational efficiency or small object detection. Future improvements in hybrid transformer architectures, optimized self-attention techniques, and real-time deployment strategies will increase the applicability of DETR for different aerial applications.

Continuous research and optimization can make DEEP learning a stronger tool in the future for autonomous aerial surveillance, environmental monitoring, and disaster response systems: a true foundation for next-generation geospatial AI models.

References:

[1] D. L. Ziyi Chen, Huayou Wang, Xinyuan Wu, Jing Wang, Xinrui Lin, Cheng Wang, Kyle Gao, Michael Chapman, "Object detection in aerial images using DOTA dataset: A survey," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 134, p. 104208, 2024, doi: https://doi.org/10.1016/j.jag.2024.104208.

[2] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp.

770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[4] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 936–944, Nov. 2017, doi: 10.1109/CVPR.2017.106.

[5] L. S. D. Peng Zhou, Xintong Han, Vlad I. Morariu, "Learning Rich Features for Image Manipulation Detection," *arXiv:1805.04953*, 2018, doi: https://doi.org/10.48550/arXiv.1805.04953.

[6] X. Zhu, W. Su, L. Lu, "Deformable DETR: Deformable Trans- formers for End- to-End Object Detection," *arXiv:2010.04159*, 2021, doi: https://doi.org/10.48550/arXiv.2010.04159.

[7] N. H. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv Prepr. arXiv2010.11929*, 2020, doi: https://doi.org/10.48550/arXiv.2010.11929.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 15979–15988, 2022, doi: 10.1109/CVPR52688.2022.01553.

[9] L. Wang and A. Tien, "Aerial Image Object Detection With Vi- sion Transformer Detector (ViTDet)," *MITRE Corp. McLean,VA,USA*, 2023, doi: https://doi.org/10.48550/arXiv.2301.12058.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12346 LNCS, pp. 213–229, 2020, doi: 10.1007/978-3-030-58452-8_13.

[11] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, Accessed: Nov. 15, 2023. [Online]. Available: https://arxiv.org/abs/1804.02767v1

[12] X. Chen, H. Fang, T. Wang, "PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection," *Int. Conf. Learn. Represent.*, 2021.

Copyright © by the authors and 50Sea. This work is licensed under the Creative Commons Attribution 4.0 International License.