

Aircraft Detection from Remote Sensing Images Using Mask R-CNN and Region of Interest Based Learning

Açmal Şüheda Semih^{1,#}, Karahalilöz Ali², Gazel Necmeddin², Sardoğan Enishan², Şentürk Melek², Sevinç Hacer Kübra¹, Karas İsmail Rakıp²

¹ Vocational School of Information Technologies, Karabuk University, Karabuk, Türkiye,
semihacmali@karabuk.edu.tr, kubrasevinc@karabuk.edu.tr

² Computer Engineering, Karabük University, Karabuk, Türkiye,
2210213009@ogrenci.karabuk.edu.tr, 2110213559@ogrenci.karabuk.edu.tr,
2210213004@ogrenci.karabuk.edu.tr, meleksenturk@karabuk.edu.tr, ismail.karas@karabuk.edu.tr

Abstract

With the proliferation of high-resolution satellite and aerial images, the need for automated systems for object detection and segmentation has emerged. Aircraft detection is a vital part of civil aviation monitoring, military intelligence, and logistical assessment. We present a comprehensive evaluation of three state-of-the-art deep learning architectures, Mask R-CNN, YOLOv8, and RetinaNet, applied to the Airbus Aircraft Detection dataset. We center our investigation on the effects of Region of Interest (RoI) based learning on the accuracy of aircraft identification and edge prediction. Our findings demonstrate a distinct trade-off between detection and spatial accuracy. YOLOv8 achieves the highest F1-score (0.9262) and Recall (0.9343), suggesting it has the best ability to identify nearly all visual indicators in a scene, whereas Mask R-CNN achieves the highest Mean Intersection over Union (mIoU) of 0.9048. This implies that RoI-based two-stage architectures are better at maintaining pixel-level structural detail compared to single-stage detectors. RetinaNet, using Focal Loss achieved the highest Precision (0.9241) while reducing false positives against more complex backgrounds. This work provides a framework for selecting a model based on the users' operational priorities: speed versus fidelity of instance segmentation. The implementation code is accessible via <https://www.kaggle.com/code/alikarahaliloz/airbus-master-v2>

Keywords: Remote Sensing, Aircraft Detection, Mask R-CNN, Region of Interest, Instance Segmentation, Satellite Imagery

1. Introduction

Remote sensing has developed to transition from low-resolution orbital observations to high-cadence, sub-meter resolution images. This development has initiated new areas of study regarding wildlife monitoring, urban planning, and infrastructure inspection (Aliane, 2025). In this realm of applications, detecting and monitoring aircraft in satellite imagery remains a central function of global surveillance. Automatically identifying, counting, and segmenting aircraft over large spatial extents is critical to understanding global transportation behaviors and national security (Mishra, 2023).

Historically, aircraft detection utilized either manual coding or shallow machine learning approaches to identify aircraft through handcrafted features. However, these approaches will quickly become untenable with the increasing volume of data from modern sensors. Deep Neural Networks (DNNs) have revolutionized Remote Sensing Image (RSI) analysis, showing strong performance in complex environments with multiple light levels changing cloud cover and scene backgrounds (Yang, 2025). Despite these successes, aircraft detection in RSI has distinct vulnerabilities in either the pixel footprint of the target, clustering within regions of the image (airports or sporting events), and extracting specific boundary landmarks for classification.

This analysis will provide performance perspectives on three different architectural philosophies in computer vision: the two-stage RoI-based approach (Mask R-CNN), a high-speed anchor-free one-stage approach (YOLOv8), and a one-stage, precision-based approach that utilizes Focal Loss (RetinaNet). We will assess their performance against the Airbus Aircraft Detection dataset to deconstruct how various learning mechanisms, and specifically the treatment of regions of interest, impact the model's ability to manage class imbalance and spatial complexity associated with satellite images.

Object detection in remote sensing has rapidly progressed from general-purpose algorithms to specific architectures aimed at the characteristics of aerial data. The early models of the YOLO (You Only Look Once) family addressed the challenge of balancing speed and accuracy and were found quite useful for real-time applications such as mapping fruit trees or doing industrial inspection (Yong, 2024). As the series evolved, later adaptations, such as the Coordinate Attention (CA) mechanism and the C2f module introduced in YOLOv8, extended the model's ability to extract multi-scale features and perceive complex scenes (Qiu, 2024).

Recent work has demonstrated the utility of modified YOLO architectures in precision agriculture, where modified models used certain band combinations with the multispectral data to count plants (Lin, 2023). In remote sensing, modified YOLOv5s have been proposed that are lighter so that embedded devices can maintain high mean Average Precision (mAP) (Zhang, 2023). Even though YOLO-based models are quick with their detections and have an acceptable recall, they frequently lack the fine-level segmentation being derived at the pixel level necessitated by objects with irregular shapes in high-resolution imagery.

Instance segmentation, which combines object detection with semantic segmentation, has emerged as a third option to put a finer point on things. The quintessential example of this approach is Mask R-CNN. In this approach, the first stage proposes regions, and the second stage refines and segments them (Sharma, 2022). In forestry, Mask R-CNN variants have been used to segment single trees in urban studies while duplicating a higher quality of masks compared to single-stage models (Lv, 2023). Conversely, in maritime surveillance, an example of an instance segmentation would use this technique on ships to obtain acutely depicted locations of objects as well as shape knowledge for ship scheduling. In this instance, instance segmentation was a considerably more difficult task than bounding box detection due to the extremely complex backgrounds of the sea surface (Yasir, 2023).

The challenge of "small objects" has remained as a recurring theme in the remote sensing literature. Small objects directly lead to high false-negative rates due to insufficient feature information. Some of the solutions developed in literature have included the standard use of the Feature Pyramid Networks (FPN) and global context through Transformer's context perception (Zhang, 2023). In addition, Focal Loss was recently proposed in the RetinaNet model as a mathematical way to overcome the extreme foreground-background class imbalance common in many satellite images (i.e. aircraft images), where the amount of area of the target object covers only a portion of the total area in a background (Zhang, 2024). This study builds on these advances by analyzing these methods rigorously in the specific case of aircraft detection.

2. Materials and Methodology

2.1. Airbus Aircraft Detection Sample Data Set and Data Preparation

The Airbus Aircraft Detection dataset, procured from Kaggle, served as the primary data source for this study. The dataset under consideration is composed of 103 high-resolution satellite images, with a spatial resolution of 2560×2560 pixels, and a total of 3,425 manually annotated instances of aircraft. The images represent a variety of environments, including international airports, aircraft storage areas, and remote airfields, providing a range of background conditions, including asphalt runways, concrete surfaces, grass fields, and desert terrain.

The annotations provide pixel-level segmentation masks for each aircraft, thus enabling object detection and instance segmentation tasks. This detailed annotation format enables the models to learn not only the location of aircraft but also their precise spatial boundaries within the images.

For the purposes of experimental evaluation, the dataset was divided into training and test subsets. The training subset comprised 80% of the images, while the remaining 20% were reserved for testing. Prior to the training phase, the images were subjected to preprocessing using standard normalisation procedures. Furthermore, fundamental data augmentation techniques such as horizontal flipping, rotation, and brightness adjustments were employed to enhance model generalisation under diverse environmental conditions.

2.2. Mask R-CNN and Region of Interest (RoI) Based Learning Architecture

Mask R-CNN represents the cutting edge of the "two-stage" philosophy of detection. Its architecture is based on Faster R-CNN with the addition of a third branch focused on predicting an object mask alongside classification and bounding box regression. The foundation of Mask R-CNN's functionality, revealing spatial precision and segmentation, lies in its RoI-based learning approaches.

The detection process begins with a backbone network—typically a ResNet with a Feature Pyramid Network (FPN), which creates a hierarchical set of feature maps at different scales. These feature maps are processed through a Region Proposal Network (RPN). The RPN processes the feature maps with sliding windows to propose regions (anchors) that are likely to be the presence of an object. While "one-stage" detectors classify directly on the grid, Mask R-CNN utilizes the region proposed to spend its computational power on areas with high probability.

A critical component of this architecture is the RoI Align layer. In the earlier versions of Mask R-CNN, RoI Pooling was utilized with quantization of the coordinates, which often temporally misaligned the image pixels from the feature maps. RoI Align addresses this concern with bilinear interpolation, leading to a computation of the exact value of the input features at four regularly sampled positions within each RoI bin and then aggregated. This process maintains the overall spatial orientation, which is likely a major contributor to Mask R-CNN achieving high pixel-level accuracy (mIoU), as it avoids "drift" of the masks associated with the "grid" or quantization (Chen, 2022). Once the features are aligned to the same spatial coordinates, the features are passed to the mask head, which applies small convolutional to generate a binary mask for each instance.

2.3. Comparative Models: YOLOv8 and RetinaNet Architecture

To fully compare the Mask R-CNN architecture, we selected the YOLOv8 and RetinaNet architectures to be representative of the "one-stage" architecture.

YOLOv8 (You Only Look Once version 8) is constructed for an extreme level of efficiency and fast inference. As compared to prior versions, YOLOv8 incorporates an anchor-free detection head that predicts the center of an object rather than the offset of a center point from an anchor box. This simplifies the model, making it much easier for the model to predict objects with different aspect ratios, which are prevalent on aircraft (e.g., long fuselages versus wide wingspans) (Zhong, 2024). YOLOv8's backbone uses a modified CSP (Cross Stage Partial) Darknet structure. The use of CSP optimizes gradient flow while reducing redundant calculations. For segmenting instances, YOLOv8 uses the "Proto" head, which generates a set of prototype masks for the entire image that are linearly combined per bounding box based on coefficients predicted for each bounding box. Although this process is much faster than a RoI-based approach for segmentation, this method generally lacks the detailed boundary precision available with Mask R-CNN, since it relies more on a global prototype set than on the local, region-specific features.

RetinaNet was included for its unique approach to addressing class imbalance issues. It is typical in satellite imagery for the ratio of aircraft pixels to background pixels to exceed 1:1000; therefore, in standard cross-entropy loss, the vast majority of the easily classified background examples dominate the loss calculation, and the model does not obtain opportunities to learn from the difficult to classify "foreground" examples. RetinaNet utilizes Focal Loss to reshape the standard cross-entropy loss to down-weight more loss assigned to more easily classified examples. Using this method, the model is forced to focus on hard cases, such as when the aircraft instance is partially occluded by shadow or instances have low contrast against the runway. RetinaNet utilizes a ResNet-FPN backbone similar to Mask R-CNN; however, it employs a one-stage method, which significantly reduces computational expense during the proposal phase and boasts precision as high as other two-stage methods. Since RetinaNet is fundamentally an object detection model and does not produce instance segmentation masks, a mask approximation was required for IoU-based evaluation. In this study, the predicted bounding boxes generated by RetinaNet were converted into binary masks by filling the bounding box region. These masks were then compared with the ground-truth segmentation masks provided in the Airbus Aircraft Detection dataset to compute mIoU. This approach enables a consistent comparison

with the instance segmentation outputs of Mask R-CNN while acknowledging the architectural differences between the models.

2.4. Training Configuration

The implementation of all models was conducted utilizing the PyTorch deep learning framework.

The training hyperparameters were defined as follows:

- **Epochs:** Training was performed for 30 epochs to ensure model convergence.
- **Optimizer:** AdamW optimizer was used with a weight decay of 0.0005.
- **Batch Size:** A batch size of 8 was used during training to balance computational efficiency and GPU memory usage.
- **Learning Rate:** The initial learning rate was set to 0.0001. A StepLR scheduler was applied to reduce the learning rate by a factor of 0.1 at the 30th epoch for fine-tuning.
- **Loss Function:** A multi-task loss function was optimized jointly, combining classification loss (L_{cls}), bounding box regression loss (L_{box}), and mask loss (L_{mask}).

3. Results

The models reached convergence through distinct optimization paths during the 30-epoch training cycle. As shown in Table 1, Mask R-CNN achieved the lowest Recall (0.8000), followed by RetinaNet (0.8754), while YOLOv8 recorded a higher Recall of 0.9343. These variations reflect how each architecture's specific head design interacts with the joint multi-task loss function.

The Mean Intersection over Union (mIoU) is the most important metric for assessing the quality of the predicted masks, as it assesses the overlap between the predicted segmentation and the true value, with a higher value indicating a better fit to the actual contours of the aircraft.

As seen in the data, Mask R-CNN outperformed the other models with a mIoU of 0.9048, supporting the original hypothesis that RoI-based learning, specifically with the RoI Align mechanism, outperforms the initially proposed model for products that could capture the irregular and fine-grained shapes of the aircraft. Mask R-CNN remains spatially focused, cropping the feature map and capturing only the proposed region while also keeping the spatial resolution of the base network for boundary recognition.

Table 1. Comparative Performance Metrics of YOLOv8, Mask R-CNN, and RetinaNet Models

Models	Mean IoU	Precision	Recall	F1 Score
Yolo v8	0.8447	0.9182	0.9343	0.9262
Mask R CNN	0.9048	0.8511	0.8000	0.8247
RetinaNet	0.8069	0.9241	0.8754	0.8991

The training curves in Figure 1 illustrate the convergence rate of each model regarding segmentation quality. Mask R-CNN demonstrates a rapid and stable ascent to high mIoU values, benefiting from the RoI Align mechanism, which preserves spatial orientation. YOLOv8 exhibits a more gradual convergence with higher variance during the early epochs, while RetinaNet reaches its performance plateau with a lower overall mIoU compared to the two-stage approach.

YOLOv8 achieved a quite poor mIoU of 0.8447. While YOLOv8's prototype-based segmentation is efficient, it typically comes away with what could be described as "smoothed" edges or slight misalignments, which are especially notable for smaller aircraft, where only a few pixels can cause the mIoU to drop significantly. RetinaNet, despite high detection precision, had the lowest mIoU at 0.8069. RetinaNet fundamentally is a detector, and when modifying or extending it to generate masks for

segmentation, the performance is not adjusted in the model with sufficient refinement to process densities. Therefore, the mask generated by RetinaNet is generally rectangular or less precise.

The visual results presented in Figure 2 illustrate the instance segmentation performance of the Mask R-CNN model on a representative satellite image from the Airbus Aircraft Detection dataset. The model successfully identifies multiple aircraft within the airport environment and produces segmentation masks that closely follow the geometric structure of aircraft fuselages and wings. The predicted masks demonstrate high spatial precision, particularly in densely populated airport regions where multiple aircraft are located near terminal buildings. The confidence scores associated with the detections are consistently high, indicating reliable model predictions. These qualitative results highlight the effectiveness of the RoI-based architecture in preserving spatial details and accurately delineating aircraft boundaries in high-resolution remote sensing imagery.

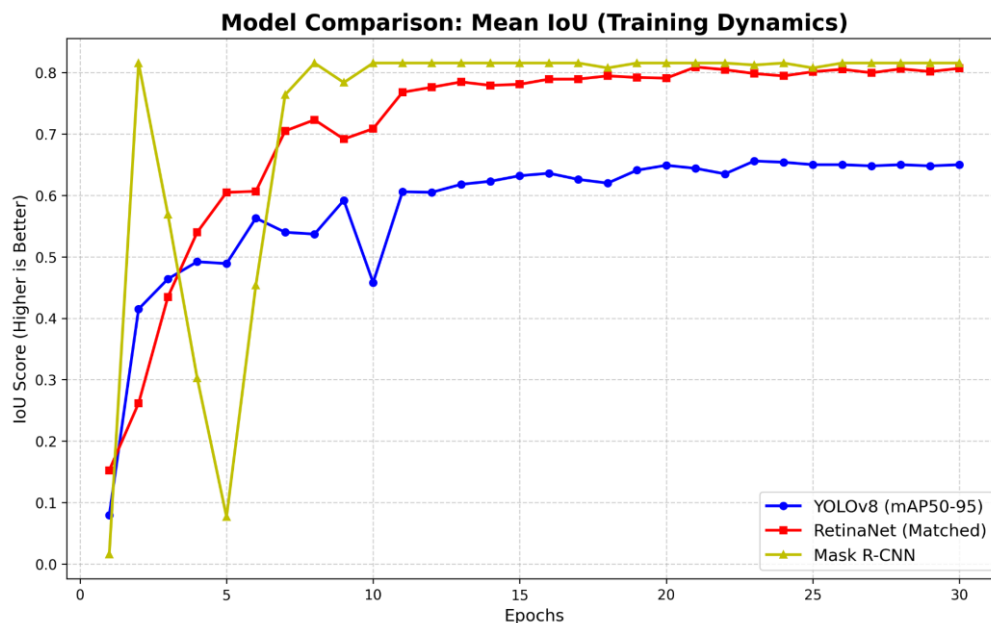


Fig. 1. Training Dynamics of Mean IoU Across Epochs for YOLOv8, Mask R-CNN, and RetinaNet

Examining the experimental data reveals that YOLOv8 is the clear winner in terms of detection reliability, achieving the highest F1 score (0.9262) and superior recall (0.9343). Essentially, the single-stage architecture in YOLOv8, when completely anchor-free, allows for the model being extremely effective at finding nearly every aircraft in the imaging, particularly with smaller or poorly lit aircraft. The recall is essential for the application of surveillance. It was also illustrated that a missed aircraft error could potentially indicate a bigger error than an inaccurately bounded aircraft.

RetinaNet achieved a Precision of 0.9241, demonstrating the effectiveness of using Focal Loss to limit false positive predictions. Therefore, RetinaNet is a good option in cluttered environments (e.g., airports, where many of the service vehicles could be mistaken for small planes). However, its Recall was lower (0.8650) than YOLOv8, suggesting it is more conservative in its predictions.

The quantitative results indicate that Mask R-CNN achieved the highest Mean IoU (0.9048) among the evaluated models, demonstrating its strong capability in capturing the precise spatial structure of aircraft instances. However, its Precision (0.8511) and Recall (0.8000) values are lower than those obtained by the one-stage detectors. This suggests that while Mask R-CNN produces highly accurate segmentation masks when an aircraft is detected, it may miss some aircraft instances compared with models such as YOLOv8 and RetinaNet. In contrast, YOLOv8 achieved the highest Recall (0.9343) and F1-score (0.9262), indicating strong detection capability in dense scenes, whereas RetinaNet obtained the highest Precision (0.9241), reflecting its effectiveness in reducing false positives. These results highlight the trade-off between segmentation accuracy and detection sensitivity across the evaluated architectures.

4. Conclusion

Our comparative analysis reveals a fundamental trade-off between detection reliability and spatial precision in remote sensing. Mask R-CNN, utilizing two-stage RoI-based learning and RoI Align, achieved the highest Mean IoU (0.9048) and pixel-level accuracy (0.9900). This confirms that local feature refinement is essential for capturing the complex geometries of aircraft. Consequently, Mask R-CNN is the optimal choice for applications requiring precise area estimation and high-fidelity instance segmentation.

Conversely, YOLOv8 emerged as the most balanced model for large-scale monitoring, achieving a superior F1-Score (0.9262) and Recall (0.9343). Its anchor-free, single-stage architecture identifies nearly every aircraft in diverse scenes, making it ideal for real-time surveillance where high detection rates are prioritized over boundary precision. RetinaNet, supported by Focal Loss, provided the highest Precision (0.9241), offering a robust solution for cluttered environments where minimizing false positives is critical.



Fig. 2. Representative Output from Mask R-CNN Segmentation Results

In conclusion, the "RoI Effect" emphasizes that specialized spatial attention mechanisms significantly improve performance against the background dominance and class imbalance of satellite imagery. While YOLOv8 excels in detection speed and reliability, Mask R-CNN remains the benchmark for spatial fidelity. Future developments should aim to integrate RoI-based refinement modules into high-speed one-stage detectors to create hybrid models capable of both rapid and pixel-perfect aircraft detection in autonomous aerial systems.

References

- Aliane, N... (2025). Drones and AI-Driven Solutions for Wildlife Monitoring. *Drones*. <https://doi.org/10.3390/drones9070455>
- Mishra, S. & Palanisamy, P... (2023). Autonomous Advanced Aerial Mobility –An End-to-end Autonomy Framework for UAVs and Beyond. <https://doi.org/10.36227/techrxiv.24523051>
- Yang, Z., Li, X., Zhang, L., & Xu, Y... (2025). Spatially Adaptive and Distillation-Enhanced Mini-Patch Attacks for Remote Sensing Image Object Detection. *Electronics*. <https://doi.org/10.3390/electronics14173433>
- Yong, X., Zeng, X., Lai, W., Liao, J., Chen, Y., Zhu, M., & Huang, K... (2024). Detecting and Mapping Individual Fruit Trees in Complex Natural Environments via UAV Remote Sensing and Optimized YOLOv5. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. <https://doi.org/10.1109/jstars.2024.3379522>
- Qiu, Y., Xue, J., Zhang, G., Hao, X., Lei, T., & Jiang, P... (2024). RS-FeatFuseNet: An Integrated Remote Sensing Object Detection Model with Enhanced Feature Extraction. *Remote Sensing*. <https://doi.org/10.3390/rs17010061>
- Lin, H., Chen, Z., Qiang, Z., Tang, S., Liu, L., & Pau, G... (2023). Automated Counting of Tobacco Plants Using Multispectral UAV Data. *Agronomy*. <https://doi.org/10.3390/agronomy13122861>
- Zhang, J., Chen, Z., Yan, G., Wang, Y., & Hu, B... (2023). Faster and Lightweight: An Improved YOLOv5 Object Detector for Remote Sensing Images. *Remote Sensing*. <https://doi.org/10.3390/rs15204974>
- Sharma, R., Saqib, M., Lin, C., & Blumenstein, M... (2022). A Survey on Object Instance Segmentation. *SN Computer Science*. <https://doi.org/10.1007/s42979-022-01407-3>
- Lv, L., Li, X., Mao, F., Zhou, L., Xuan, J., Zhao, Y., Yu, J., Song, M., Huang, L., & Du, H... (2023). A Deep Learning Network for Individual Tree Segmentation in UAV Images with a Coupled CSPNet and Attention Mechanism. *Remote Sensing*. <https://doi.org/10.3390/rs15184420>
- Yasir, M., Zhan, L., Liu, S., Wan, J., Hossain, M. S., Colak, A. T. I., Liu, M., Islam, Q. U., Mehdi, S. R., & Yang, Q... (2023). Instance segmentation ship detection based on improved Yolov7 using complex background SAR images. *Frontiers in Marine Science*. <https://doi.org/10.3389/fmars.2023.1113669>
- Zhang, Y., Han, D., & Chen, P... (2023). Swin-PAFF: A SAR Ship Detection Network with Contextual Cross-Information Fusion. *Computers, materials & continua/Computers, materials & continua (Print)*. <https://doi.org/10.32604/cmc.2023.042311>
- Zhang, M., Wang, Z., Song, W., Zhao, D., & Zhao, H... (2024). Efficient Small-Object Detection in Underwater Images Using the Enhanced YOLOv8 Network. *Applied Sciences*. <https://doi.org/10.3390/app14031095>
- Chen, W., Shang, G., Ji, A., Zhou, C., Wang, X., Xu, C., Li, Z., & Hu, K... (2022). An Overview on Visual SLAM: From Tradition to Semantic. *Remote Sensing*. <https://doi.org/10.3390/rs14133010>
- Zhong, H., Zhang, Z., Liu, H., Wu, J., & Lin, W... (2024). Individual Tree Species Identification for Complex Coniferous and Broad-Leaved Mixed Forests Based on Deep Learning Combined with UAV LiDAR Data and RGB Images. *Forests*. <https://doi.org/10.3390/f15020293>