Hue University Journal of Science: Techniques and Technology; pISSN 2588-1175 | eISSN 2615-9732

Vol. 134, No. 2B, 2025, 19–31; DOI: 10.26459/hueunijtt.v134i2B.7862

Evaluation of the effectiveness of modern object detection models on spatial image data

Dung Nguyen[1], Van-Dung Hoang[2]*, Van-Tuong-Lan Le[3]

[1] University of Sciences, Hue University, Hue, Vietnam
[2] Faculty of Information Technology, HCMC University of Technology and Education, Ho Chi Minh, Vietnam
[3] Department of Academic and Students' Affairs, Hue University, Hue, Vietnam

Abstract. Object detection in aerial imagery, especially from unmanned aerial vehicles (UAVs), presents numerous challenges due to varying altitudes, occlusions, and diverse object scales—particularly the detection of small objects. This paper provides a comparative evaluation of three advanced object detection models: YOLOv11, RT-DETR, and RF-DETR, using the VisDrone2019 dataset, which includes complex urban and suburban scenes captured from UAVs. We analyze the models based on key performance metrics such as mean average precision (mAP), inference speed, model size, and computational complexity. Experimental results show that YOLOv11 achieves the highest processing speed, making it especially suitable for real-time applications due to its fast inference and strong edge-device performance. RF-DETR, on the other hand, achieves the best accuracy, with the fastest mAP@0.5 and mAP@[0.5:0.95] scores of 46.9% and 26.6%, respectively, demonstrating effectiveness in complex scenarios with high object density and occlusions. RT-DETR offers a balanced trade-off between speed and accuracy, making it a practical choice for applications requiring both responsiveness and reliable detection quality. These findings clarify the strengths and limitations of each model and provide practical guidance for selecting suitable object detection models in UAV-based surveillance and tracking tasks.

Keywords: Transformer, Computer Vision, Object detection, Detection Transformer

1 Introduction

Object detection is one of the core problems in computer vision, with a wide range of practical applications such as security surveillance [1], intelligent transportation systems [2], autonomous vehicles [3], and especially aerial surveillance using Unmanned Aerial Vehicles (UAVs) [4]. In recent years, the rapid advancement of deep learning models has significantly enhanced the performance of object detection systems, particularly in complex real-world scenarios.

However, object detection from UAV imagery remains a major challenge due to characteristics such as oblique viewing angles, varying altitudes, small object sizes, occlusions, and uneven lighting conditions [5]. To address these challenges, many advanced models have been proposed to improve spatial feature extraction and contextual correlation modeling.

* *Corresponding: nguyendung@hueuni.edu.vn*

Received: 14–05–2025; Accepted: 25–06–2025

Among modern detection models, the YOLO (You Only Look Once) family [6-16] is notable for its high processing speed and real-time deployment capability, while Transformer-based models such as DETR (Detection Transformer) [17-19] and its variants have demonstrated superior accuracy by modeling global relationships in images. Recently, models like YOLOv11 [16], RT-DETR (Real-Time DETR) [20], and RF-DETR (Roboflow DETR) [21] have achieved state-of-the-art (SOTA) results on various benchmark datasets.

In addition, the VisDrone2019 dataset is one of the most comprehensive and challenging UAV vision datasets to date, featuring diverse urban and traffic scenes captured under varying weather conditions and camera perspectives. Thus, VisDrone2019 is a suitable benchmark for evaluating the generalization and robustness of modern object detection models in realistic conditions.

The main contributions of this paper are as follows:

- A comprehensive comparative evaluation of three SOTA object detection models, YOLOv11, RT-DETR, and RF-DETR, on the VisDrone2019 dataset.
- Performance analysis of the models based on multiple evaluation metrics, including accuracy (mAP), inference speed (FPS), computational complexity (GFLOPs), and model size.
- Discussion of the strengths and limitations of each model in specific scenarios, providing practical recommendations for selecting suitable models in UAV-based surveillance and object tracking applications.

2 Related Work

2.1 Object Detection from UAV Imagery

In recent years, the use of unmanned aerial vehicles (UAVs) to capture aerial imagery has become increasingly common across various domains, including traffic monitoring, precision agriculture, and urban management. However, the unique characteristics of UAV imagery, such as high resolution, small object sizes, oblique viewing angles, and varying lighting conditions, pose significant challenges for object detection tasks, especially when compared to ground-level images or videos.

A major contribution that has driven research in UAV-based object detection is the VisDrone dataset, introduced by Zhu et al. (2018) [22]. This dataset contains over 10,000 manually annotated frames collected from 14 cities across China, featuring diverse object categories such as pedestrians, cars, buses, and trucks. In the benchmark evaluation, popular models like Faster R-CNN [23], SSD [24], RetinaNet [25], and YOLOv3 [8] were tested, but their performance remained limited. For instance, a variant of RetinaNet, HAL-RetinaNet [22], achieved only 46.18%

mAP@0.5 and 31.88% mAP on the VisDrone2018-DET test set [22], highlighting the complexity of UAV data and the urgent need for improving existing detection models.

Subsequent research has focused on enhancing feature extraction capabilities and designing network architectures better suited to UAV data. In [26], Zhang et al. (2021) proposed a multi-scale adversarial network where features are learned jointly from UAV and satellite images using a Siamese architecture. This approach improved the model's adaptability to challenging lighting conditions and complex scenes and outperformed conventional CNN-based models on both UAVDT [27] and VisDrone [28] datasets, achieving up to a 10.3% increase in mAP compared to baseline methods like YOLOv3 and Faster R-CNN.

Additionally, Xia et al. (2024), in a study published on SSRN [29], introduced the PSPPF module combined with the EFFN network to effectively exploit multi-scale information in UAV images. This architecture is specifically designed for detecting small objects such as motorcycles and cars from high altitudes. It emphasizes parameter reduction and inference acceleration, while also improving detection accuracy through efficient pooling mechanisms and specialized detection heads. Experimental results on the VisDrone2019 dataset [28] show that the method achieved 46.2% mAP@50 and 28.0% mAP@[0.5:0.95], demonstrating strong performance in detecting small-scale vehicles in UAV imagery. Although it lacks direct quantitative comparison with SOTA models like YOLO, the method shows promising efficiency and practical applicability.

Overall, these studies highlight current trends in UAV object detection research. Notably, deep learning models with multi-scale feature extraction and fusion mechanisms are increasingly adopted to improve detection of small and indistinct objects in complex environments. Another important direction is the design of lightweight modules that optimize parameter count and inference speed to meet the real-time deployment requirements on UAV platforms. These trends not only address current technical challenges but also pave the way for broader application of UAV-based image analysis and surveillance systems in the future.

2.2 Advanced Object Detection Methods

In recent years, modern object detection models have made significant advancements in both accuracy and processing speed, largely driven by the emergence of Transformer architectures and hybrid models that combine Convolutional Neural Networks (CNNs) with Transformers. Among these, three models—YOLOv11, RT-DETR, and RF-DETR—have demonstrated outstanding performance on various benchmark datasets.

YOLOv11 is one of the most recent and advanced versions in the YOLO object detection series. It inherits the strengths of previous versions (from YOLOv5 to YOLOv8) and introduces several key architectural innovations to improve both detection accuracy and inference speed. The architecture of YOLOv11 consists of three major components:

- Backbone: YOLOv11 adopts a novel hybrid backbone that incorporates the C3k2 module, a lightweight and efficient convolutional block designed to reduce computational cost while preserving strong feature representation. The backbone also integrates SPPF (Spatial Pyramid Pooling - Fast) to enhance multi-scale context aggregation, and C2PSA, a spatial attention mechanism that helps the model focus on salient features, especially useful for small and occluded objects.

- Neck: The Neck module consists of multiple stacked C3k2 blocks with up-sampling and feature concatenation operations. This structure facilitates effective feature fusion across different scales, enabling the model to better detect objects of various sizes. Attention mechanisms (C2PSA) are also applied at this stage to further refine feature maps.

- Head: The detection head is built using CBS (Convolution–BatchNorm–SiLU) and C3k2 blocks, followed by a Detect layer. It supports anchor-free detection and is designed to efficiently localize and classify objects with high precision, especially in real-time scenarios.

YOLOv11 supports multiple computer vision tasks beyond object detection, including instance segmentation, pose estimation, and oriented object detection (OBB), making it highly versatile. Moreover, it is available in various model sizes (from nano to extra-large), allowing deployment across a range of devices, from edge hardware to high-end GPUs. The overall architecture of YOLOv11 is illustrated in Fig. 1.

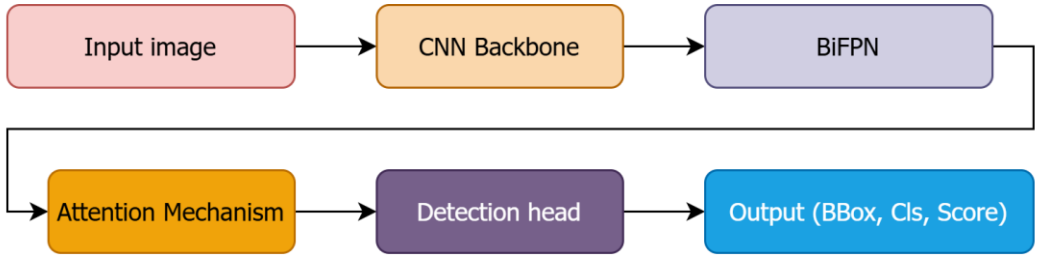


Fig. 1. The overall architecture of YOLOv11

In contrast, **RT-DETR** is a real-time object detection model developed by Baidu, designed to enhance the inference speed and accuracy of the original DETR architecture. RT-DETR optimizes convergence and reduces computational cost through the use of an efficient hybrid encoder. The model consists of two main components:

- **AIFI (Attention-based Intra-scale Feature Interaction):** Integrates features within the same scale to reduce computational overhead while preserving detection accuracy.

- **CCFM (Cross-Scale Cross-Feature Fusion):** Fuses features across multiple scales to improve the detection of small objects without significantly increasing computational cost.

In addition, RT-DETR employs an **IoU-aware query selection** mechanism, which helps select more accurate object queries from the outset and eliminates the need for Non-Maximum

Suppression (NMS). As a result, RT-DETR achieves near real-time inference speeds on common GPUs while maintaining high mean Average Precision (mAP) scores on datasets such as COCO [30]. The overall architecture of RT-DETR is illustrated in Fig. 2.

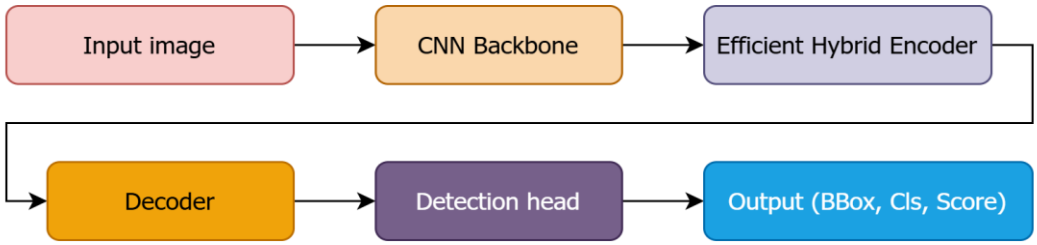


Fig. 2. The overall architecture of the RT-DETR model

RF-DETR is a real-time object detection model developed by Roboflow, designed to tackle practical detection challenges requiring fast and efficient processing. It employs a DINOv2 Backbone alongside a simplified Transformer architecture (inspired by Deformable DETR and LW-DETR) to strike a balance between performance and computational complexity.

- **No NMS Approach:** A key innovation in RF-DETR is the elimination of Non-Maximum Suppression (NMS). Instead, it relies on intelligent query selection and IoU-aware optimization mechanisms, enabling end-to-end inference that is both accurate and fast.

- **Performance on COCO:** RF-DETR-Large achieves **60.5 mAP@[0.5:0.95]** at a 728-pixel input resolution, marking it as the first real-time model to surpass the 60 mAP threshold on COCO—processing at approximately 25 FPS (≈6 ms/image) on an NVIDIA T4 GPU.

- **Model Variants:** The model is offered in two sizes—**Base** (29 M parameters, ~53.3 mAP) and **Large** (128 M, 60.5 mAP)—catering to both edge devices and high-performance systems.

- **Domain Adaptability:** Beyond COCO, RF-DETR excels on Roboflow's **RF100-VL** benchmark, achieving ~86.7 mAP@50, demonstrating strong generalizability to diverse real-world domains

The overall architecture of RF-DETR is illustrated in Fig. 3. It features:

- **DINOv2 Transformer Backbone** – pre-trained for strong global context modeling and cross-domain generalization.

- **Simplified Transformer Encoder-Decoder** – with deformable cross-attention layers tailored for single-scale feature extraction, reducing inference overhead.

- **Efficient Querying** – IoU-aware query filtering eliminates post-processing via NMS.

- **Multi-resolution Training Support** – enables flexible latency-accuracy trade-offs without retraining.

This architecture enables RF-DETR to deliver real-time inference (~6 ms/image) without compromising on accuracy, making it suitable for applications like security surveillance, autonomous driving, and industrial automation. The lightweight Base version further enables deployment on edge devices with limited compute while maintaining strong detection performance.

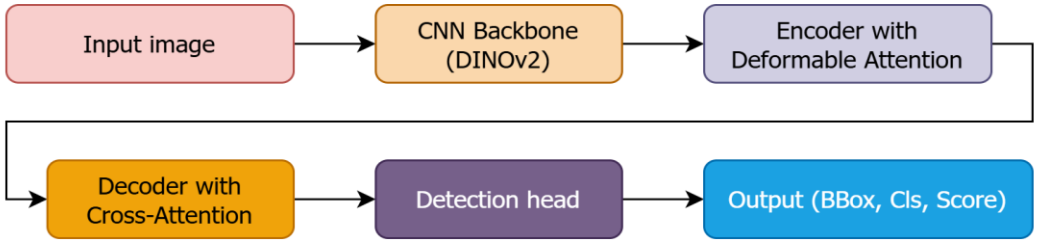


Fig. 3. The overall architecture of the RF-DETR model

Overall, the models discussed above reflect a clear trend towards transitioning from traditional CNN architectures to Transformer-based or hybrid designs, with a strong emphasis on real-world deployment. This includes prioritizing inference speed, accuracy, and domain flexibility. **Table 1** below provides a detailed comparison of the technical specifications and performance metrics of the three models—YOLOv11, RT-DETR, and RF-DETR, all of which were trained on the COCO [30] dataset.

Table 1. Comparison of model specifications and performance (Trained on COCO Dataset)

Attribute	YOLOv11	RT-DETR	RF-DETR
Main Architecture	CNN + Transformer (hybrid)	Transformer Decoder + FPN	Transformer + DINOv2 backbone
NMS	Yes	No	No
Typical Backbone	CSPDarknet [31]	ResNet [32]	DINOv2 [33]
Transformer Decoder	Yes	Yes (multi-layer)	Yes (simplified)
Number of Parameters	~2.6M–56.9M	~32M–76M	~32M–128M
FPS (on GPU T4/V100)	~88.5–666.7 FPS	~66.5–198.8 FPS	~166.7 FPS
mAP	~39.5–54.7	~53.0–54.8	Up to 60.5 (RF-DETR-Large, 728px)
Inference Time per Image	~1.5–11.3 ms	~5.03–15.03 ms	~6 ms (Base, on GPU T4)

Attribute	YOLOv11	RT-DETR	RF-DETR
Model Complexity GFLOPs	6.5–194.9	136–259	N/A
Edge Device Deployment	Good (with small version)	Possible (depending on backbone)	Very good (Base for edge)
Release	Ultralytics, 2024	Baidu, 2023	Roboflow, 2025

3 Dataset

This study utilizes the VisDrone2019 dataset [28], one of the leading benchmark datasets for object detection in images captured from UAVs. The dataset was developed by the AISKYEYE group at Tianjin University, China, and was published as part of international computer vision competitions. VisDrone2019 includes over 10,000 static images extracted from 263 UAV-captured videos in 14 different cities across China, under diverse environmental, weather, and viewing angle conditions. Each image in the dataset is manually annotated with detailed information, including bounding boxes and object types. There are 10 common object classes in urban and traffic environments, including: pedestrians, cyclists, motorcyclists, bicycles, motorcycles, tricycles, cars, small trucks, large trucks, and buses. The dataset is divided into three parts: training set (6,471 images), validation set (548 images), and test set (3,190 images). The images in VisDrone vary in resolution, from 960×540 to over 2000×1500 pixels, reflecting changes in the UAV's viewing angles and flight heights.

VisDrone is considered one of the most challenging datasets due to its high ratio of small objects, significant occlusions, complex scenes, and the requirement for models to balance accuracy and processing speed. During evaluation, models are tested based on strict criteria consistent with COCO standards, including mAP at multiple IoU thresholds (from 0.5 to 0.95), mAP@0.5. The VisDrone dataset is widely used as a key benchmark in object detection research from UAVs and serves as a basis for comparing the performance of modern models. Fig. 4 illustrates some sample images from the dataset.

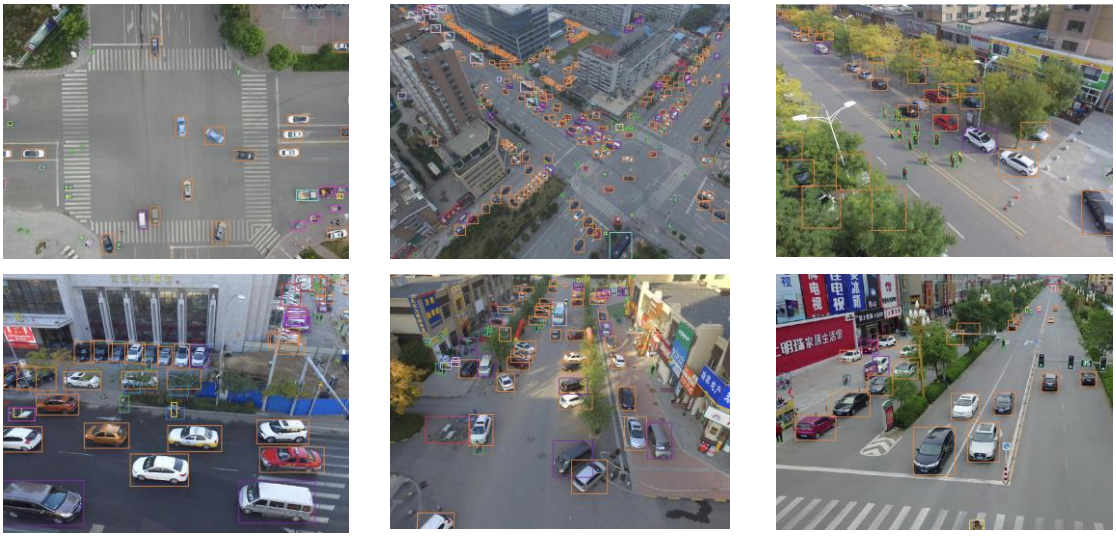


Fig. 4. Some images from the dataset

4 Experiments

4.1 System Specifications and Training Settings

The experiments were conducted on a high-performance computing system with the following specifications, and the models were trained using pre-trained weights from ImageNet (for the backbone) and COCO (for object detection), then fine-tuned on the VisDrone2019 dataset. Details of the system specifications and training hyperparameters are provided in Table 2.

Table 2. System specifications and training hyperparameters

Category	Specification
GPU	NVIDIA GeForce RTX 4090
CPU	13th Gen Intel(R) Core(TM) i7-13700F
RAM	64 GB
Operating System	Ubuntu 22.04.4 LTS
Framework	TensorFlow 2.4.0+cu121
Batch Size	4
Learning Rate	0.001
Optimizer	AdamW
Epochs	100

4.2 Evaluation Metrics

The performance of the models was evaluated using the following metrics:

- **mAP@0.5**: Mean average precision at an intersection over union (IoU) threshold of 0.5
- **mAP@0.5:0.95**: Mean average precision averaged from IoU 0.5 to 0.95 with a step size of 0.05
- **Number of Parameters (Parameters)**: Total learnable parameters of the model
- **FLOPs**: Floating-point operations, indicating the computational complexity

These metrics help evaluate both the accuracy and real-time performance of the models.

4.3 Experimental Results

After training and evaluating the YOLOv11, RT-DETR, and RF-DETR models on the VisDrone2019 dataset, we obtained the results as presented in Table 3. The metrics used for evaluation include detection accuracy (mAP), the number of parameters, FLOPs, and inference speed (FPS), to provide a comprehensive view of the performance of each model.

Table 3. Performance Comparison of Models on VisDrone2019

Model	mAP@0.5	mAP@[0.5:0.95]	Parameters	FLOPs
YOLOv11-n	33.5%	19.6%	2.6M	6.5G
RT-DETR-l	37.3%	22.0%	32.8M	108G
RF-DETR-Base	46.9%	26.6%	32.0M	60G

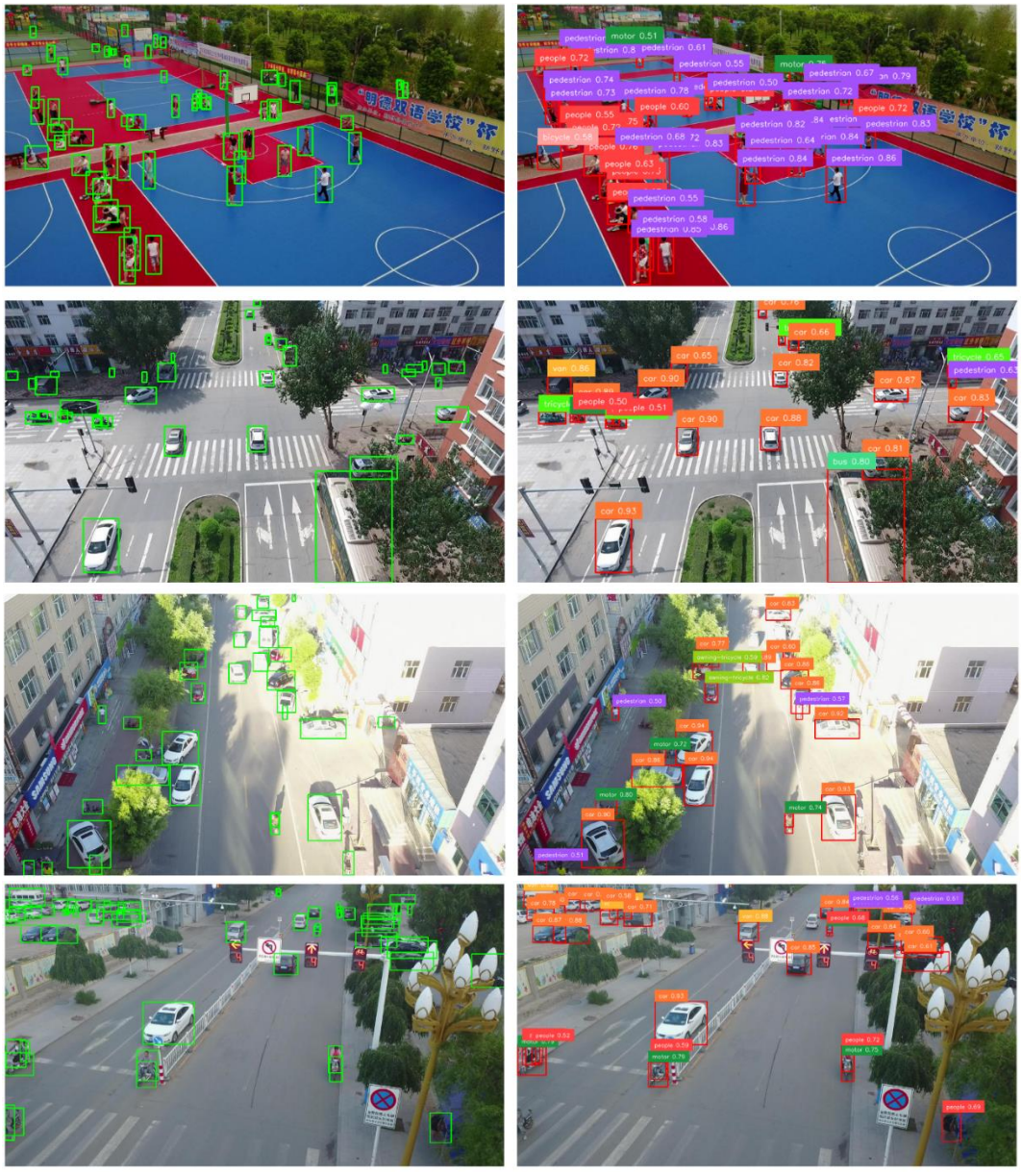
Regarding accuracy, RF-DETR achieves the highest performance with a mAP@0.5 of 46.9% and mAP@[0.5:0.95] of 26.6%, demonstrating superior detection capabilities in complex scenarios, particularly due to its region-focused attention mechanism. RT-DETR comes second with a mAP@[0.5:0.95] of 22.0%, while YOLOv11 shows lower accuracy but still proves effective in object detection with minimal computational cost.

In terms of inference speed, YOLOv11 stands out with a speed of around 125 FPS, significantly higher than RT-DETR (85 FPS) and RF-DETR (69 FPS). This comes from YOLOv11's lightweight design, with only 2.6 million parameters and 6.5 GFLOPs, allowing efficient deployment on UAV devices with limited resources.

Overall, YOLOv11 is the most suitable model for real-time applications due to its balance between speed, sufficient accuracy, and lightweight deployment capability. On the other hand,

RT-DETR and RF-DETR are better choices for scenarios requiring higher accuracy or working in environments with dense and complex object distributions.

Figure 5 shows some images with the Ground Truth and Predicted Bounding Boxes, along with the class names and confidence of each object.



(a) Ground Truth *(b) Predict*

Fig. 5. Some comparison images between Ground Truth and Predictions

5 Conclusion

Object detection in UAV images remains a challenging problem due to factors such as altitude, occlusion, and object size diversity. Through experimental evaluation on the VisDrone2019 dataset, this paper compared three advanced detection models: YOLOv11, RT-DETR, and RF-DETR. The results show that each model has its own strengths: YOLOv11 is suitable for real-time applications due to its fast inference speed; RF-DETR excels in accuracy in complex scenarios; while RT-DETR achieves a good balance between speed and accuracy. These analyses provide a comprehensive view of the performance of each model, aiding researchers and practitioners in selecting the most suitable solution for surveillance and object tracking tasks from UAVs in real-world conditions. Future research could focus on combining the strengths of current models to enhance the detection of small objects in complex environments while ensuring real-time processing speed.

Acknowledgements

We would like to thank the University of Hue for sponsoring this research under the grant number DHH2025-01-226.

References

1. S. M. Gilani, A. Anjum, A. Khan, M. H. Syed, S. A. Moqurrab, and G. Srivastava, "A robust Internet of Drones security surveillance communication network based on IOTA," *Internet of Things*, vol. 25, p. 101066, 2024.
2. M. Bakirci, "Utilizing YOLOv8 for enhanced traffic monitoring in intelligent transportation systems (ITS) applications," *Digital signal processing*, vol. 152, p. 104594, 2024.
3. H. Wang, C. Liu, Y. Cai, L. Chen, and Y. Li, "YOLOv8-QSD: An improved small object detection algorithm for autonomous vehicles based on YOLOv8," *IEEE Transactions on Instrumentation and Measurement*, 2024.
4. M. N. Mowla, D. Asadi, K. N. Tekeoglu, S. Masum, and K. Rabie, "UAVs-FFDB: A high-resolution dataset for advancing forest fire detection and monitoring using unmanned aerial vehicles (UAVs)," *Data in brief*, vol. 55, p. 110706, 2024.
5. A. A. Laghari, A. K. Jumani, R. A. Laghari, H. Li, S. Karim, and A. A. Khan, "Unmanned aerial vehicles advances in object detection and communication security review," *Cognitive Robotics*, 2024.
6. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779-788.
7. J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263-7271.
8. J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

9. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
10. G. Jocher *et al.*, "ultralytics/yolov5: v3. 0," *Zenodo*, 2020.
11. C. Li *et al.*, "Yolov6 v3. 0: A full-scale reloading," *arXiv preprint arXiv:2301.05586*, 2023.
12. C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464-7475.
13. G. J. a. A. C. a. J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.
14. C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv preprint arXiv:2402.13616*, 2024.
15. A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, no. 107984-108011, 2024.
16. R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.
17. A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
18. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020: Springer, pp. 213-229.
19. D. Nguyen, V.-D. Hoang, and V.-T.-L. Le, "V-DETR: Pure Transformer for End-to-End Object Detection," in *Asian Conference on Intelligent Information and Database Systems*, 2024: Springer, pp. 120-131.
20. Y. Zhao *et al.*, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16965-16974.
21. I. a. R. Robinson, Peter and Popov, Matvei, "RF-DETR," *https://github.com/roboflow/rf-detr*, 2025.
22. P. Zhu *et al.*, "Visdrone-det2018: The vision meets drone object detection in image challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0-0.
23. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
24. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016: Springer, pp. 21-37.
25. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.
26. S. N. Ruiqian Zhang, Zhenfeng Shao, Xiao Huang, Jiaming Wang, Deren Li, "Multi-scale adversarial network for vehicle detection in UAV imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021.
27. D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370-386.

28. D. Du *et al.*, "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0-0.

29. H. Xia, Y. Liu, and W. Li, "Vehicle detection in UAV aerial imagery based on multi-scale feature fusion," *Available at SSRN 5042527*.

30. T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 2014: Springer, pp. 740-755.

31. C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390-391.

32. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

33. M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.