УДК 004.93:004.8

HYBRID CNN-TRANSFORMER MODEL-BASED OBJECT DETECTION METHODOLOGY FOR UAV IMAGERY

Nguyen Van Bach, bach253@gmail.com
Belarusian State University of Informatics and Radioelectronics

Abstract. This paper proposes a hybrid object detection method for UAV imagery by combining YOLO (CNN-based) and RT-DETR (transformer-based) models. The framework integrates a bounding box processing module with fusion strategies and a confidence-based selection mechanism. Experimental results show that the hybrid approach improves detection accuracy, achieving a 0.7 increase in mAP50.

Keywords: UAV, object detection, YOLO, RT-DETR, hybrid model.

Introduction. Unmanned aerial vehicle (UAV) applications impose stringent demands on object detectors, requiring both rapid inference and high accuracy across diverse perspectives and flight altitudes. At present, to address the problems, convolutional neural network (CNN)–based models (e.g., You Only Look Only (YOLO)) [1] and transformer-based models (e.g., Real-Time DEtection TRansformer (RT-DETR)) [2] have been used. Cons of CNN-based detectors are reduced precision for small objects, difficulty with dense scenes and occlusion, localization errors with objects of unusual shapes or sizes, and limited versatility. In turn, the cons of transformer-based detectors are high computational cost, slow convergence, difficulties with small object detection, and sensitivity to initialization and training conditions.

This project presents an object detection framework for UAV imagery that can combine multiple detection modules to provide the trade-off between detection robustness, accuracy and real-time performance. In the project, we propose an integrated approach based on using a CNN-based object detector (YOLO) and a transformer-based object detector model (a Real-Time DETR (RT-DETR)) (Figure 1). Integrating CNN-based and transformer-based object detectors offers a powerful hybrid approach that combines the strengths of both architectures for more robust and accurate object detection, especially in complex scenes where understanding the relationships between objects is important, and with small objects. First detector excels at local feature extraction, while the second detector excels at capturing long-range dependencies and global context. However, this approach provides enhanced accuracy, improved small object detection, global context awareness, and flexibility at the expense of increasing computational cost, requirements for training data, and difficulties with optimizing architecture and training strategy for a hybrid CNN-transformer model. A hybrid CNN-transformer model is implemented by the proposed modules: a hybrid neural network model-based detection module and a bounding box processing and analysis module.

Integration of a CNN-based detector (YOLO) and a lightweight transformer-based detector (RT-DETR) to generate candidate bounding boxes is based on using a confidence score that reflects how likely the module output is correct. Higher scores mean the module is more confident in its prediction, while lower scores suggest more uncertainty. The output of these modules contains multiple candidate object bounding boxes in the format XYXY. The bounding box processing and analysis module consists of submodules based on bounding box fusion rules and confidence-based bounding box selection. The bounding box fusion rule-based submodule allows us to compute the set of IoU values for the predicted candidate object bounding boxes in format XYXY to select bounding boxes related to the same classes and compute new bounding boxes for them by means of non-maximum suppression (NMS), soft-NMS, weighted box fusion (WBF) [3] or non-maximum weighted (NMW) rules [4]. To reduce false positive boxes, we use a confidence-based selection submodule that computes the IoU between two boxes belonging to different classes and compares it to the IoU-threshold value.

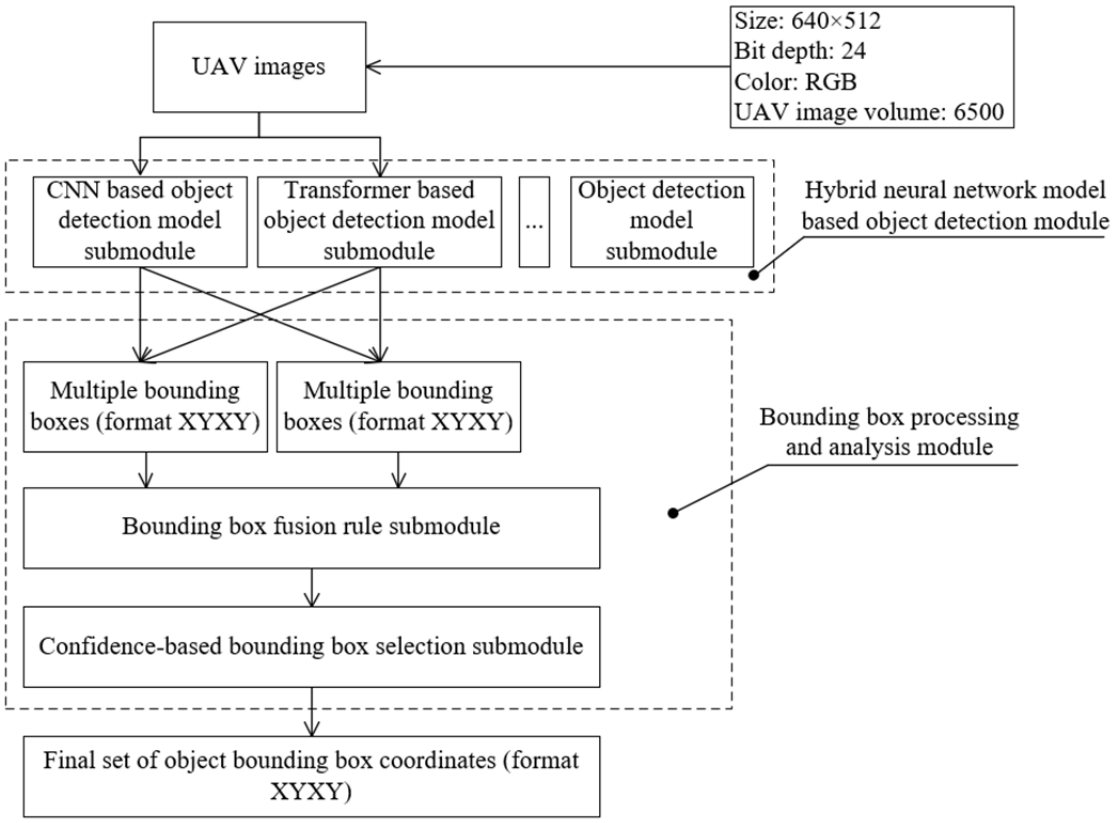

Figure 1. – Integrated modular object detection structure

Experiment. Figure 2, **Figure 3** and **Figure 4** show that the mean average precision values of object detection for small (yellow), medium (blue), large (green) object size and total mean average precision values (violet) obtained by using YOLOv10-based submodule, RT-DETR-based submodule, and hybrid module with using WBF submodules, respectively. These models were transfer-learned on a training set of 5200 UAV images and evaluated on a validation set of 1300 images for vehicle detection.

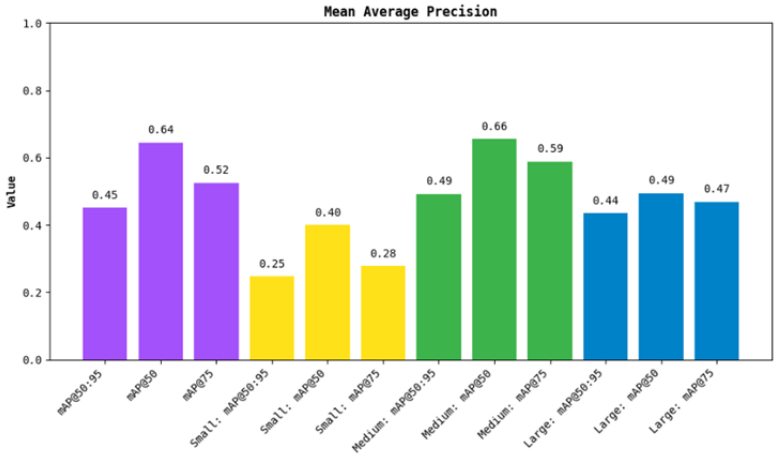

Figure 2. – Mean average precision values of object detection for the YOLOv10-based submodule

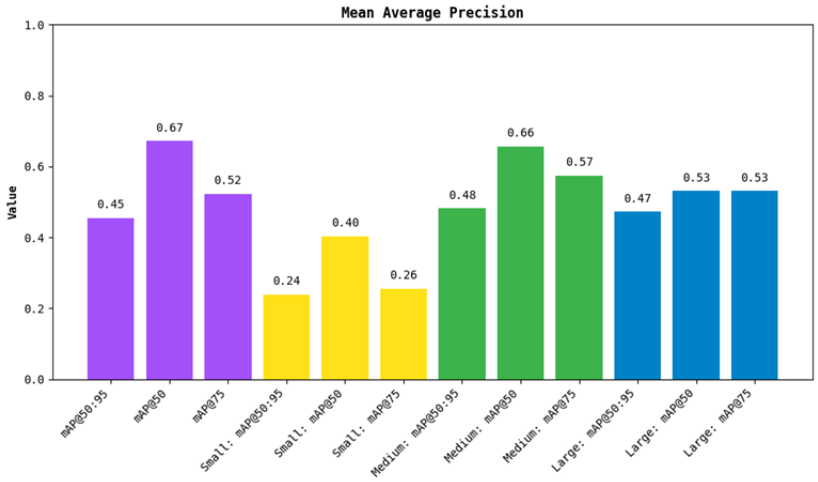

Figure 3. – Mean average precision values of object detection for RT-DETR-based submodule

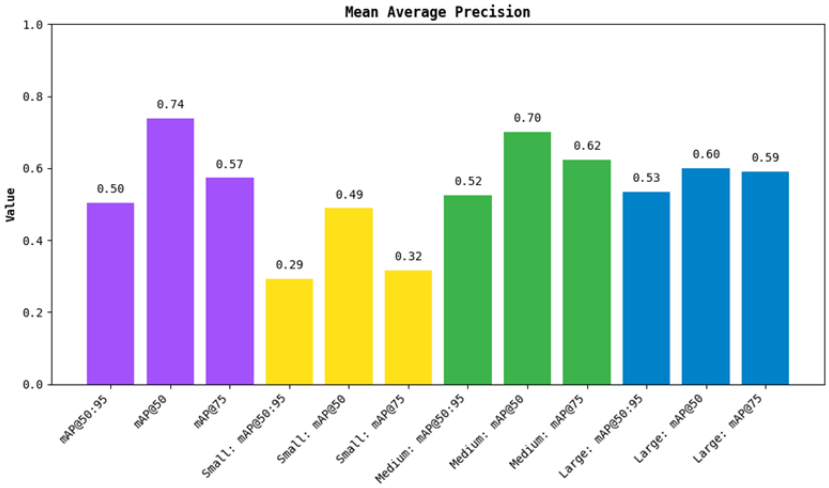

Figure 4. – Mean average precision values of object detection using the hybrid module

It follows from Figure 2, Figure 3 and Figure 4 that the mean precision value of the detection object is increased by 0.7 mAP_{50} by using the hybrid module for the given UAV imagery.

The Figure 5 illustrates the detected objects obtained using the proposed hybrid module.

Figure 5. – Object detection results using the proposed hybrid module

Conclusion. A hybrid CNN-transformer framework for UAV object detection was presented. By combining YOLO and RT-DETR and applying bounding box fusion and confidence-based selection, the method improves detection accuracy and reduces false positives.

References

1. Wang, A. Yolov10: Real-time end-to-end object detection / H. Chen, L. Liu, K. Chen, Z. Lin, & J. Han // Advances in Neural Information Processing Systems. 2024. – P. 107984-10801.

2. Zhao, Y. Detrs beat yolos on real-time object detection / W. Lv, S. Xu, J. Wei // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2024. – P. 16965-16974.

3. Solovyev, R. Weighted boxes fusion: Ensembling boxes from different object detection models / W. Wang, & T. Gabruseva, // Image and Vision Computing. – 2021. – Vol. 107, – P. 104117.

4. Zhou, H. Cad: Scale invariant framework for real-time object detection / Z. Li, C. Ning, & J. Tang // Proceedings of the IEEE international conference on computer vision workshops. – 2017. – P. 760-768.