The CNN-Transformers Crossroads, Comparing RT-DETR and YOLOv12 for Small object detection in remote sensing images

Behnam Solatinia *, Saeid Niazmardi, Tayeb Alipour Fard

Department of Surveying Engineering, Faculty of Civil and Surveying Engineering, Graduate University of Advanced Technology, Kerman, Iran – b.solatinia @student.kgut.ac.ir, (s.niazmardi, t.alipour)@kgut.ac.ir))

KEY WORDS: Object Detection, Small Object Detection, Remote sensing, Deep Learning, RT-DETR, YOLOv12, Hybrid Architecture, Performance Evaluation

ABSTRACT:

Detecting small objects in remote sensing images has always been a challenge. The Convolutional Neural Network (CNN) and Transformer-based networks are two prominent categories of deep learning models used to address this challenge. Recently, combining both architectures has emerged to improve detection performance. However, a direct comparison between the leading standard models representing these architectures has yet to be conducted. In this study, we provided a performance comparison of two state-of-the-art detectors: YOLOv12, a CNN-based model with an attention mechanism, and RT-DETR, a transformer-based model built on a CNN backbone. We fine-tuned both algorithms on a custom remote sensing dataset containing small objects (airplanes and cars) and evaluated their performance based on precision, recall, F1-score, and training time. The results showed that YOLOv12 was significantly faster to train and achieved higher precision. These qualities make it a better choice for applications where minimizing false positives is critical. RT-DETR, with high recall and F1-score, was more effective at detecting a larger number of small objects. This analysis offers valuable insights into the trade-offs between these two architectures and serves as a guideline for selecting the appropriate model for each specific remote sensing task.

1. INTRODUCTION

Object detection is a crucial task in computer vision that involves identifying and locating objects in an image or video (Zaidi et al., 2022). Object detection in remote sensing images has various applications across different field domains, such as, land use and land cover monitoring (Ait El Haj et al., 2023; Kanagasundaram et al., 2022; Taiwo et al., 2023), traffic management (Ahmadi et al., 2019; Kopsiaftis & Karantzalos, 2015; Macioszek & Kurek, 2021), urban planning (Xue & Zhao, 2022), disaster management (Teodoro & Duarte, 2022), environmental monitoring (Cheng & Dang, 2022), and object tracking (Chen et al., 2022; Chou et al., 2024; Zhang et al., 2022).

Given the diverse applications of object detection, various detection algorithms have been proposed. Among which, deep learning methods are the most favored ones. Considering the used architectures, deep object detectors can be categorized into two main categories: Convolutional Neural Network (CNN)-based and transformer-based detectors.

Most of the earlier object detection algorithms proposed in the computer vision society, such as Faster Regions with Convolutional Neural Network (Faster R-CNN), Single Shot Multi-Box Detector (SSD), and You Only Look Once (YOLO) (Liu et al., 2016; Redmon et al., 2016; Ren et al., 2017) are CNN-based networks that rely on convolutional layers as their backbone for feature extraction.

Recently, transformer-based networks have been applied in several vision tasks including object detection. The DEtection TRansformer (DETR) and its variants such as Real-time DEtection TRansformer (RT-DETR) and Arbitrary-oriented object detection transformer (AO2-DETR) are examples of transformer-based detectors with outstanding results (Carion et al., 2020; Dai et al., 2022; Zhao et al., 2024).

Algorithms from both categories have been widely employed as off-the-shelf solutions in the remote sensing literature. However, the objects in remote sensing images are usually small, which challenges the performance of these detectors. Most detectors cannot detect small objects or falsely detect noise as objects. To address this issue, researchers have modified algorithms from both categories to enhance small-object detection.

Several CNN-based algorithms have been modified for detecting small objects in remote sensing images. For example, an improved Faster R-CNN method was introduced for detecting small ships in high-resolution images (Zhang et al., 2019). This method incorporates a specific convolutional modification to the VGG16 backbone of Faster R-CNN, which produces a multi-resolution feature map and performs ROI pooling on a larger feature map within the Region Proposal Network (RPN). The results demonstrated high accuracy and recall. To improve the detection accuracy and speed of small object detection, a modified version of YOLOV3, named YOLO-fine was proposed (Pham et al., 2020). This algorithm simplifies the Darknet-53 backbone by removing its last two convolutional blocks, which contain several parameters to balance accuracy and training speed.

Several modified transformer-based detectors have also been proposed. For example, in (H. Zhang et al., 2024), the authors proposed the RS-DETR, a modified version of RT-DETR. This detector enhances the attention-driven feature interaction (AIFI) module by integrating it with Cascaded Group Attention (CGA) while maintaining the CNN backbone the same as RT-DETR. In 2024, Drone-DETR was proposed as an enhancement of RT-DETR to improve the accuracy of small object detection (Kong et al., 2024). Their method introduces the Effective Small Object Detection Network (ESDNet) as a lightweight backbone to increase the accuracy of small object detection and an Enhanced

* Corresponding author

Dual-Path Feature Fusion Attention Module (EDF-FAM) for the neck network to improve performance for multi-scale object detection.

Recent studies focused on merging CNNs and transformers to improve the accuracy of detectors. These two architectures can be combined using various scenarios. Here, we have focused on YOLOv12 and RT-DETR, which are two state-of-the-art models in the computer vision society. YOLOv12 is the latest version of the YOLO family, a CNN-based architecture that now uses an attention mechanism to improve feature extraction. RT-DETR is a leading DETR variant that is designed for real-time performance and combines a CNN backbone for efficient feature extraction with a transformer-based decoder for object queries.

Both models are powerful solutions that incorporate the hybrid CNN-transformer approach and are available as off-the-shelf models. Nevertheless, their performance for small object detection in remote sensing images has not been evaluated. Thus, this study conducts a performance comparison of these two models for the specific task of small object detection in remote sensing. We fine-tune and evaluate YOLOv12 and RT-DETR on a challenging custom dataset to analyze their strengths and weaknesses in terms of precision, recall, and F1-score. Also, their training time was compared to assess their efficiency. This evaluation will help the researchers select the most suitable model for remote sensing applications where accuracy and speed are essential.

2. METHOD

2.1 YOLOv12

The YOLO algorithm was introduced in 2016 as a real-time and end-to-end model for object detection in the field of computer vision (Redmon et al., 2016). Unlike other CNN-based detectors that adopt separate stages for region proposal, feature extraction, classification, and bounding box regression, YOLO unifies these stages into a single, fully convolutional network. This design simplifies the detection pipeline, allowing fast training, inference, and easy deployment.

Since 2016, 12 versions of YOLO have been developed. Each version enhanced the architecture in terms of speed, accuracy, robustness, and usability. These improvements, combined with the algorithm's ability to achieve high performance after fine-tuning with small datasets, make the YOLO algorithm one of the most popular object detection models in both research and real-world applications.

YOLOv12 is the latest YOLO model that aims to balance the inference speed and object detection accuracy by architectural improvements. Unlike earlier YOLO versions that used Efficient Layer Aggregation Network (ELAN), this version's backbone uses a Residual Efficient Layer Aggregation Network (R-ELAN), which improves feature aggregation by merging deep convolutional layers with residual connections, and improves its robustness. This modification helps the model to improve gradient flow and optimization issues, and also enhances multi-scale feature extraction.

Another key improvement of the YOLOv12 algorithm over previous YOLO models is its architecture, which is built around attention mechanisms. YOLOv12 model adopts an Area Attention (A2) module to focus on the important regions of the feature map while keeping low computation costs.

The A2 module divides the feature map into horizontal or vertical segments and then uses an efficient attention mechanism known as FlashAttention to minimize memory access overhead. This enables YOLOv12 to utilize attention context without incurring a high computational cost. Also, YOLOv12 uses powerful lightweight enhancement convolution to increase speed and accuracy, such as 7×7 separable convolutions, which replace positional encodings and enable spatial awareness with fewer parameters (Tian et al., 2025). These modifications to the architecture of the YOLO model make the new version faster than many of its older versions and improve object detection accuracy in challenging situations, such as small objects and noisy scenes.

Figure 1 shows the architecture of YOLOv12, which has convolutional layers and R-ELAN enhanced by the Area Attention and a lightweight convolution block (A2C2f). The network processes input images of 640 × 640 pixels, which are reduce through a series of convolutional and residual aggregation block to extract multi-scale features. The backbone begins with an initial convolution layer (3×3 kernel, stride 2, 64 channels), followed by additional convolutional layers and C3k2 Bottleneck block that reduce spatial resolution from 640×640 to 320×320, 160×160, 80×80, 40×40, and 20×20 while extending feature depth up to 1024 channels. The model utilizes an anchor-free detection head, eliminating the need for predefined anchor boxes and allowing more adaptive bounding box regression. (Jegham et al., 2025).

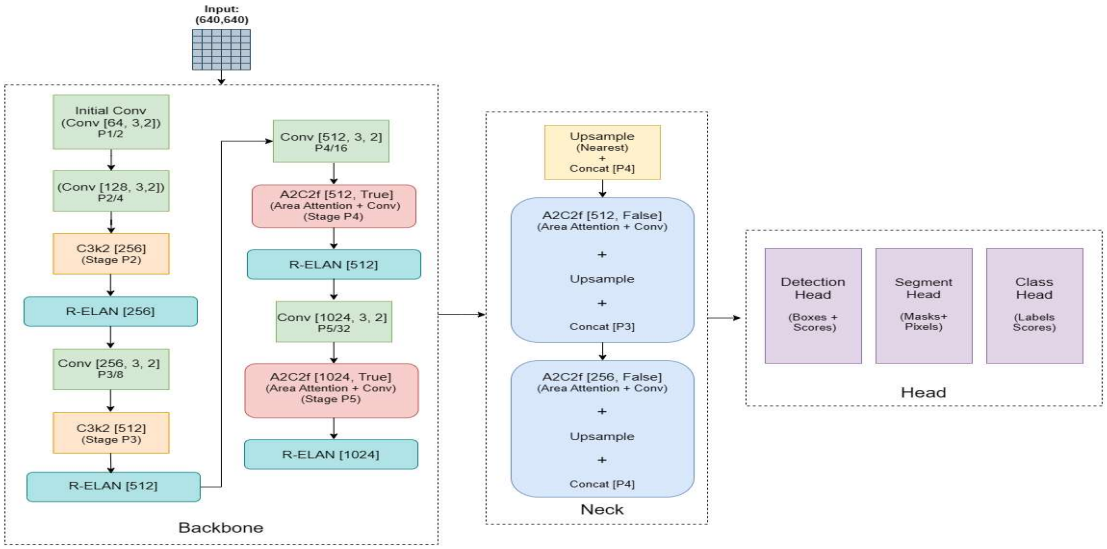


Figure 1. YOLOv12 architecture (Jegham et al., 2025).

2.2 RT-DETR

DETR is a detection algorithm that eliminates the need for components like non-maximum suppression (NMS), anchor generation, and proposal stages in previous detectors (Carion et al., 2020).

Nevertheless, the DETR algorithm faces several challenges, including the high computational cost of its standard transformer encoder and slow training. To address these challenges while keeping its key benefits, RT-DETR was introduced (Zhao et al., 2024). RT-DETR introduces several modifications to DETR architecture for high detection accuracy and faster inference. One of its main modifications is adopting a lightweight transformer decoder combined with a convolutional backbone, such as ResNet or CSPDarkNet, for visual feature extraction. Besides, RT-DETR designs object queries based on their localization confidence, which helps the model minimize the inclusion of low-quality queries that negatively affect the accuracy of object localization and class prediction (Zhao et al., 2024). These modifications improve the model's robustness and training time, which leads to improved performance on small object detection in complex scenes.

Figure 2 shows the RT-DETR architecture. The architectural diagram shows the final three stages of the backbone (S3, S4, S5) as input to the encoder, which extracts multi-scale feature maps from the CNN network. The encoder converts these features into sequential representations, and give it to the efficient hybrid encoder, which integrates the Adaptive Interaction Feature Integration (AIFI) and the Cross-scale Complementary Feature Fusion (CCFF) modules. AIFI module improves spatial and semantic feature representation by enabling adaptive interaction among multi-scale features, while CCFF module fuses information to preserve both fine-grained details and global context. Also, RT-DETR utilizes a 6 layer encoder-decoder transformer, with 300 object queries, which selectively updates only high-confidence queries at each iteration to reducing computational cost (Zhao et al., 2024).

3. DATASET AND EXPERIMENTS

3.1 Data

For fine-tuning, and evaluating algorithms, we have used a custom data set using images sourced from the VISO data set (Yin et al., 2021) and Roboflow website. The VISO data set is a satellite video data set focused on small object tracking. The combined data set includes 113 remotely sensed images, annotated for two object categories of airplanes and cars. The data set was split into the train, and validation subsets, containing 93, and 20 images respectively. Three images have been chosen for testing the algorithm's performance.

Figure 3, shows the test images with the ground truth and the number of objects for each image tabulated in Table 1. These images were chosen to test algorithms in challenging situations, such as small-size, low-contrast, and multi-scale objects in remote sensing images.

Image	Airplanes	Cars
a	1	38
b	3	20
c	1	73

Table 1 Number of objects includes in test images

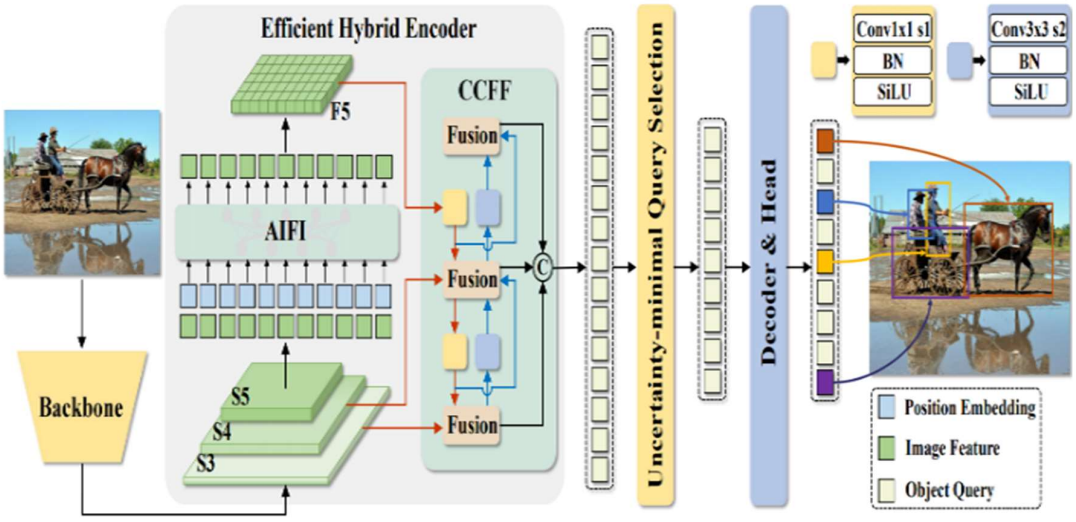
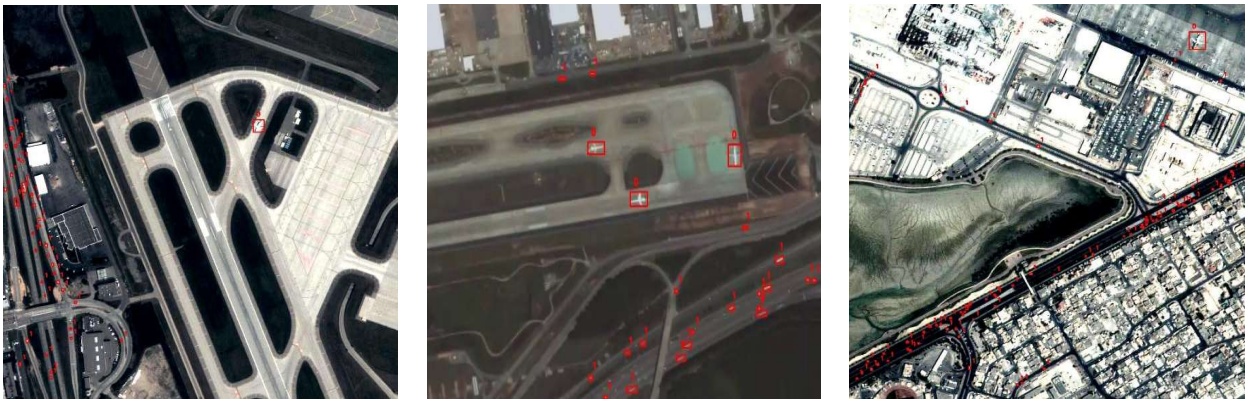


Figure 2. Overview of RT-DETR architecture (Zhao et al., 2024).



(a) (b) (c)

Figure 3. The test images for evaluating the performance of algorithms.

3.2 Implementation

To evaluate and compare the object detection performance and training times for small object detection, YOLOv12 and RT-DETR were fine-tuned on the introduced data set. To this end, pre-trained weights were initialized to leverage previously learned low-level features and improve convergence on the small remote sensing data set.

To ensure a fair comparison between the training times and performance of algorithms, we have selected the weights versions with the most similar number of layers and hyperparameters. Weights information and training settings tabulated in table 2.

	YOLOv12	RT-DETR
Layers	169	302
Hyperparameters (million)	20106454	31987850
Optimizer	AdamW	AdamW
GFLOPs	67.1	103.4
Input image size	640 × 640	640 × 640

Table 2. Algorithm's information

We used the Ultralytics package (Glenn Jocher et al., 2023) on the Google Colab framework, which runs Tesla T4 16GB GPU and Python 3 terminal, to fine-tune algorithms. Both algorithms were trained for 100 epochs, and the patience number was set to 20 to stop the training process if no changes were observed during the last 20 epochs. Also, 70% of the GPU was used for training, which scored 11-12 GB for each epoch.

F1-Score, precision, and recall were used for accuracy metrics. Precision measures false detections, higher precision indicates fewer false positives. Recall, crucial for identifying missed objects, reflects the model's ability to correctly detect objects. F1-Score introduced as the harmonic mean of precision and recall that balances them in a single metric (Taha & Hanbury, 2015). Equations 1 – 3 denote the formula for each metrics.

$$Precision = \frac{True\ Positives}{True\ Positives + Fals\ \ Positives} \quad (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

4. RESULTS

Training times of detection algorithms are tabulated in Table 3.

Algorithm	Time (minutes)
YOLOv12	12.12
RT-DETR	18.48

Table 3. Training time

Due to the small data set, pre-trained weights were used. Even though RT-DETR stopped early in the 96th epoch due to meeting the patience number, it required a longer time than YOLOv12. The higher speed of YOLO in the training phase can be attributed to its architectural design, which is optimized for speed, and its fewer parameters (20M in YOLOv12 compared to 33M in RT-DETR). Because of attention mechanisms and object query matching, RT-DETR contains 302 layers compared to YOLOv12's 169 layers and requires more processing for training. Also, YOLOv12 needs fewer GFLOPs, which helps the model perform faster with fewer computations.

To compare the detection performance of the algorithms, precision, and recall for each object class are tabulated in Table 4.

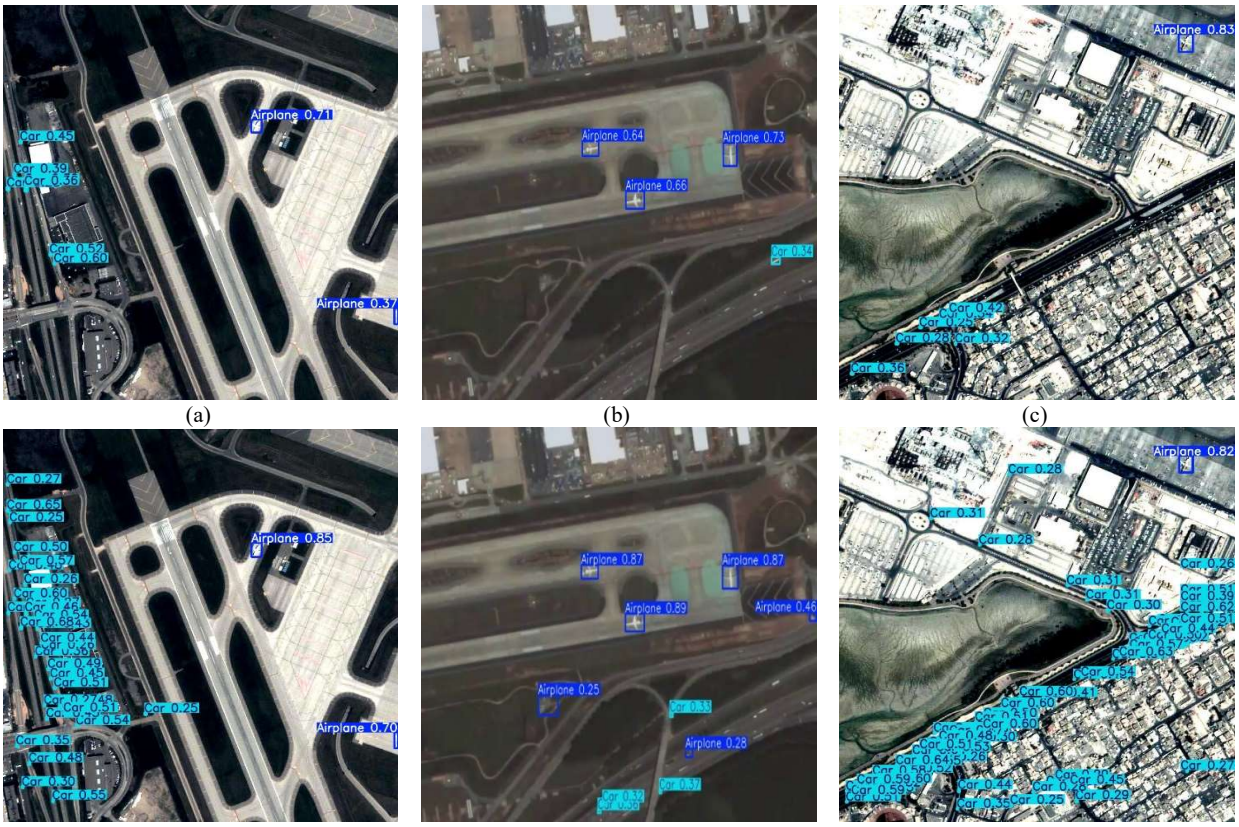
Image	Class	Index	YOLOv12	RT-DETR
a	Airplane	Precision	50	50
	Car	Precision	71.43	51.61
		Recall	13.16	42.11
b	Airplane	Precision	100	50
		Recall	100	100
	Car	Precision	100	25
		Recall	5	5
c	Airplane	Precision	100	50
		Recall	100	100
	Car	Precision	50	28.77
		Recall	4.11	28.77

Table 4. Object detection accuracy for each class

Airplanes, due to their specific morphology, were easier to detect, and both algorithms achieved 100% recall. In terms of precision, YOLOv12 performed better, achieving the highest value in images b and c (100%), while RT-DETR, due to its false detections, reached 50% across all the images. Additionally, RT-DETR detected an airplane in image c twice due to its query-based prediction without NMS, which helps the model reduce duplicate detections when multiple queries focus on the same object.

Detecting cars in images is more challenging because of their small size and large quantity. YOLOv12 achieved acceptable precisions in all images, however, its recall is very low at 13.16%, 5%, and 4.11% for images a, b, and c, respectively. These values for recall indicate that it missed a large number of cars in each image. RT-DETR's lower precision and higher recall show its ability to detect more small objects.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W8-2025-721-2026 | © Author(s) 2026. CC BY 4.0 License.



(a) (b) (c)

(d) (e) (f)

Figure 4. visualization object detection results, object detection using YOLOv12 (a, b, c), and RT-DETR (d, e, f)

Table 5 lists the obtained accuracies from both algorithms in terms of precision, recall, and F1-score.

Algorithm	Index	Images		
		a	b	c
YOLOv12	Precision	66.67	100	57.14
	Recall	15.38	17.39	5.41
	F1-Score	25	29.63	9.88
RT-DETR	Precision	51.52	40	29.33
	Recall	43.59	17.39	29.73
	F1-Score	47.22	24.24	29.53

Table 5. Object detection accuracy on the test images

In terms of precision, YOLOv12 outperformed RT-DETR across all three images. The most notable difference appears in Image (b), where YOLOv12 achieves 100% compared to 40% for RT-DETR.

These results showed that YOLOv12 focuses on high-confidence prediction, which helps reduce false detections but causes it to miss many objects.

RT-DETR scored the highest recall in images with a greater number of small objects. For Image (c), YOLOv12 yielded a low recall of 5.41%, while RT-DETR yielded 29.73%, which shows its advantage in detecting small objects. This higher accuracy is due to its transformer-based attention mechanism of RT-DETR.

Considering the F1-score, for images a and c, RT-DETR performs better, highlighting its ability to detect small objects. However, in Image (b), YOLOv12 scored a higher F1-score due to its perfect precision on that image.

Figure 4 shows the object detection results for test images, and the differences among the number of detected objects in the class of car by each algorithm are obvious.

5. CONCLUSION

Detecting small objects in remote sensing has always been a challenge. Recently, the hybrid CNN-Transformer object detectors have shown outstanding results. In this study, we selected an attention-enhanced CNN model (YOLOv12) to a hybrid CNN-transformer model (RT-DETR) and compared to their ability in detecting small objects in remote sensing images. The results highlighted a clear trade-off among the two architectures. YOLOv12, as a convolution-based model optimized for speed, was faster to train and achieved higher precision across all test images. This makes it a strong choice for applications where minimizing false detections is important. However, its strength in high-confidence predictions leads to lower recall, as it failed to detect a significant number of small objects.

RT-DETR, as a transformer-based model, illustrated a higher ability to detect small objects, resulting in higher recall and more balanced F1-scores in the images with a high number of objects. Its attention mechanism and query-based detection allow it to

detect more small objects, making it a better choice for applications such as traffic monitoring or environmental surveillance, where detecting all objects is necessary.

Our findings fit into a broader body of research focused on creating new hybrid architectures. While we evaluated two leading standard models, other studies have proposed custom solutions by modifying existing models, such as FFCA-YOLO, which added feature enhancement and fusion modules to YOLO for small objects (Y. Zhang et al., 2024), LAR-YOLOv8, which improved YOLOv8's feature extraction by using a dual-branch attention mechanism and a vision transformer block (Yi et al., 2024), and YOLOv5, which was modified by combining it with a Swin Transformer backbone and a weighted bidirectional feature pyramid network in the neck (Cao et al., 2023).

Comparing YOLOv12 and RT-DETR provides a valuable baseline, which shows even standard models present acceptable performance. This suggests that while specialized models are acceptable, there is still significant reason to design architectures that better balance precision and recall for remote sensing tasks.

We should consider that our evaluation was done on single-frame images. Future work could explore the ability of models in satellite videos to evaluate their performance and robustness over time. Also, research could focus on creating new architectures that combine the high speed and precision of YOLO's convolutional model with the higher detection completeness of RT-DETR's transformer.

REFERENCES

Ahmadi, S. A., Arsalan, G., Mohammadzadeh, A., 2019: Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: a perspective on a smarter city. *International Journal of Remote Sensing*, *40*(22), 8379-8394. https://doi.org/10.1080/01431161.2019.1610983

Ait El Haj, F., Ouadif, L., Akhssas, A., 2023: Monitoring land use and land cover changes using remote sensing techniques and the precipitation-vegetation indexes in Morocco. *Ecological Engineering & Environmental Technology*, *24*.

Cao, X., Zhang, Y., Lang, S., Gong, Y., 2023: Swin-Transformer-Based YOLOv5 for Small-Object Detection in Remote Sensing Images. *Sensors*, *23*(7).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020: End-to-end object detection with transformers. European conference on computer vision.

Chen, R., Ferreira, V. G., Li, X., 2022: Detecting Moving Vehicles from Satellite-Based Videos by Tracklet Feature Classification. *Remote Sensing*, *15*(1), 34.

Cheng, Q., Dang, C. N., 2022: Using GIS Remote Sensing Image Data for Wetland Monitoring and Environmental Simulation. *Computational Intelligence and Neuroscience*, *2022*, 7886358. https://doi.org/10.1155/2022/7886358

Chou, Y. S., Lee, P. J., Bui, T. A., Hsu, P. H., 2024, 6-8 Jan. 2024: Enhanced Moving Object Detection and Tracking in Remote Sensing Videos. 2024 IEEE International Conference on Consumer Electronics (ICCE).

Dai, L., Liu, H., Tang, H., Wu, Z., Song, P., 2022: AO2-DETR: Arbitrary-oriented object detection transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(5), 2342-2356.

Glenn Jocher, Qiu Jing, Chaurasia, A., 2023: *Ultralytics YOLO*. https://ultralytics.com

Jegham, N., Koh, C. Y., Abdelatti, M., Hendawi, A., 2025: YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions. *arXiv preprint arXiv:2411.00201*. https://arxiv.org/abs/2411.00201

Kanagasundaram, G., Dissanayake, K., Samarasuriya, C. 2022, 27-29 July 2022: Remote Sensing and GIS Approach to Monitor the Land-Use and Land-Cover Change in Kaduwela Metropolitan Area. 2022 Moratuwa Engineering Research Conference (MERCon).

Kong, Y., Shang, X., Jia, S., 2024: Drone-DETR: Efficient Small Object Detection for Remote Sensing Image Using Enhanced RT-DETR Model. *Sensors*, *24*(17), 5496. https://www.mdpi.com/1424-8220/24/17/5496

Kopsiaftis, G., Karantzalos, K., 2015, 26-31 July 2015: Vehicle detection and traffic density monitoring from very high resolution satellite video data. 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., 2016: SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, M. Welling, *Computer Vision – ECCV 2016* Cham.

Macioszek, E., Kurek, A., 2021: Extracting Road Traffic Volume in the City before and during covid-19 through Video Remote Sensing. *Remote Sensing*, *13*(12).

Pham, M.-T., Courtrai, L., Friguet, C., Lefèvre, S., Baussard, A., 2020: YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images. *Remote Sensing*, *12*(15), 2501. https://www.mdpi.com/2072-4292/12/15/2501

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016, 27-30 June 2016: You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ren, S., He, K., Girshick, R., Sun, J., 2017: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Taha, A. A., Hanbury, A., 2015: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, *15*, 1-28.

Taiwo, B. E., Kafy, A. A., Samuel, A. A., Rahaman, Z. A., Ayowole, O. E., Shahrier, M., Duti, B. M., Rahman, M. T., Peter, O. T., Abosede, O. O., 2023: Monitoring and predicting the influences of land use/land cover change on cropland characteristics and drought severity using remote sensing techniques. *Environmental and Sustainability Indicators*, *18*, 100248. https://doi.org/https://doi.org/10.1016/j.indic.2023.100248

Teodoro, A. C., Duarte, L., 2022: Chapter 10 - The role of satellite remote sensing in natural disaster management. In A. Denizli, M. S. Alencar, T. A. Nguyen, D. E. Motaung (Eds.),

Nanotechnology-Based Smart Remote Sensing Networks for Disaster Prevention (pp. 189-216). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-323-91166-5.00015-X

Tian, Y., Ye, Q., Doermann, D., 2025: Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.

Xue, T., Zhao, X., 2022: Dynamic monitoring of urban planning based on image data fusion in multi-source remote sensing. In *Advances in Geology and Resources Exploration* (pp. 494-502). CRC Press.

Yi, H., Liu, B., Zhao, B., Liu, E., 2024: Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 1734-1747. https://doi.org/10.1109/JSTARS.2023.3339235

Yin, Q., Hu, Q., Liu, H., Zhang, F., Wang, Y., Lin, Z., An, W., Guo, Y., 2021: Detecting and tracking small and dense moving objects in satellite videos: A benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-18.

Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M., Lee, B., 2022: A survey of modern deep learning based object detection models. *Digital Signal Processing*, *126*, 103514. https://doi.org/https://doi.org/10.1016/j.dsp.2022.103514

Zhang, H., Ma, Z., Li, X., 2024: RS-DETR: An Improved Remote Sensing Object Detection Model Based on RT-DETR. *Applied Sciences*, *14*(22), 10331. https://www.mdpi.com/2076-3417/14/22/10331

Zhang, J., Jia, X., Hu, J., Tan, K., 2022: Moving Vehicle Detection for Remote Sensing Video Surveillance With Nonstationary Satellite Platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5185-5198. https://doi.org/10.1109/TPAMI.2021.3066696

Zhang, S., Wu, R., Xu, K., Wang, J., Sun, W., 2019: R-CNN-Based Ship Detection from High Resolution Remote Sensing Imagery. *Remote Sensing*, *11*(6), 631. https://www.mdpi.com/2072-4292/11/6/631

Zhang, Y., Ye, M., Zhu, G., Liu, Y., Guo, P., Yan, J., 2024: FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1-15. https://doi.org/10.1109/TGRS.2024.3363057

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2024: Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,