Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1736319

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 1: 287-294 / January 2026

EVALUATION OF CUTTING-EDGE OBJECT DETECTION ARCHITECTURES ON MULTI-OBJECT AND SINGLE-OBJECT DATASETS

Cevahir PARLAK[1]*

[1]*Fenerbahçe University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 34758, İstanbul, Türkiye*

Abstract: This study focuses on the performance evaluation of cutting-edge object detection models, namely, YOLO12X, Mask R-CNN, RT-DETR-X, and RF-DETR-Large on the Open Images (Multi-Object) and LaSOT (Single-Object) datasets. Current cutting-edge trend applications involve CNN-based and Transformer-based object detection models. CNN-based models can use one-pass (YOLO family) or two-pass (R-CNN family) implementations. One-pass object detection models can be faster but suffer from accuracy compared to the two-pass models. Transformer-based models can use Detection Transformers or Vision Transformers. Transformer-based models are gaining popularity, and their performance surpasses CNN-based models. This study evaluates YOLO12X, Mask R-CNN from CNN-based family, and RT-DETR-X, RF-DETR-Large transformer-based architectures in terms of accuracy and time on the Open Images and the LaSOT datasets. All models are the largest available models and pretrained on COCO dataset. Transformer-based models incorporate special types of self-attention and pose significant improvement both on accuracy and speed. The experimental results demonstrate that attention and transformer-based models perform better than the traditional CNN-based object detectors and YOLO12X is the fastest method with a far margin. On the LaSOT dataset, RT-DETR-X posts 0.8804 IoU, 0.7047 F1-score, 0.6597 mAP@0.5, 28.64 fps whereas YOLO12X achieves 0.8572 IoU, 0.6657 F1-score, 0.5357 mAP@0.5, and 49.78 fps.

Keywords: Object detection, Transformers, Convolutional neural networks

***Corresponding author:** Fenerbahçe University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, 34758, İstanbul, Türkiye
E mail: cevahir.parlak@fbu.edu.tr (C. PARLAK)
Cevahir PARLAK https://orcid.org/0000-0002-5500-7379

Received: July 07, 2025
Accepted: December 22, 2025
Published: January 15, 2026

Cite as: Parlak, C. (2026). Evaluation of cutting-edge object detection architectures on multi-object and single-object datasets. *Black Sea Journal of Engineering and Science*, *9*(1), 287-294.

1. Introduction

Object detection is a formidable and outstanding key computer vision task and demonstrated remarkable developments throughout the last half century. Early studies started with Template Matching (Yuille, 1991) using predefined templates to match the images. In 1990's, feature-based methods such as Haar Cascades (Viola and Jones, 2001), SIFT (Scale Invariant Feature Transform) (Lowe, 2004), and HOG (Histogram of Oriented Gradients) (Dalal and Triggs, 2005) took over the image processing field. After the advent of deep learning. CNN-based models started to dominate the field including the R-CNN family (Girshick et al., 2014). R-CNN, Fast R-CNN (Girshick et al., 2015), and Faster R-CNN (Ren et al., 2015) improved the object detection performance significantly. SSD (Single-Shot Detector) (Liu et al., 2016) was another approach using multi-scale feature maps to overcome the performance issues of R-CNN family. YOLO (You Only Look Once) (Redmon et al., 2016) was proposed to provide a solution for real-time object detection applications giving a trade-off between the accuracy and speed. The performance of YOLO is usually behind R-CNN-based models, but it is very fast compared to the R-CNN family detectors. Finally, attention and transformer-based models are used in object detection and provide high accuracy while reducing the computational overhead. Mask R-CNN (He et al., 2017) detectors are going beyond simple bounding box detection by detecting the precise mask of the objects. Today, two mainstream object detectors are competing: CNN-based models and Transformer-based models. CNN-based object detectors can be implemented as one-stage and two stage. One-stage detectors scan the image only once and two-stage detectors scan the image two times. The YOLO series and SSD (Single-Shot Detectors) are examples of one-stage detectors and therefore they are faster than two-stage R-CNN family detectors. Transformer-based models use DETR (Detection Transformer) and ViT (Vision Transformer) (Dosovitskiy et al., 2021) architectures. DETR architectures eliminate the need of Non-Maximum Suppression used in CNN-based models which speeds up the process. ViT models, on the other hand, divide the image into patches, usually 16, and each patch is used as an input vector (Carion et al., 2020; Chen et al., 2024; Sun et al., 2024).

Object detection has very wide application areas including self-driving vehicles, medical imaging, security, surveillance, smart cities, industrial production, quality control, agriculture, environmental surveillance, multimedia, and social media. It will enable machines to understand and interpret what they see. Object detection is very essential in autonomous vehicles to track objects such as pedestrians, traffic signs, other cars, and obstacles on the road. Object detection can also be used for autonomous drones to recognize possible targets and take immediate necessary action.

The studies on object detection are extremely widespread and steadily expanding. Atik et al. (2022) compared different versions of YOLO on aerial images of DOTA dataset (Xia et al., 2018) using recall, precision, and F1-score metrics. Their data has 43 images including 9 different object classes. YOLOv2 (Redmon and Farhadi, 2017) demonstrated a better performance however, YOLOv3 (Redmon and Farhadi, 2018) has significant advantage on speed.

Ereken and Tarhan (2025) studied on the Mask R-CNN for weapon detection. Their model is based on FPN (Feature Pyramid Network) and Resnet101 classifier and pretrained on COCO (Lin et al., 2014) dataset. The model reached 0.81 mAP on the training set, 0.78 mAP precision on the validation and test sets on a 700-image dataset.

Bakır et al. (2023) studied the YOLOv5 (Jocher, 2020) model on noisy environments. In their experiments, YOLOv5 first was trained on a high-quality image dataset and then tested on the Military Aircraft dataset with different Gaussian noise levels. YOLOv5 reached up to 0.735 mAP on the clean dataset, however, on the noisy datasets, the mAP scores were 0.486, 0.356, 0.244, and 0.235 for the 30%, 50%, 80%, and 100% noisy data respectively.

Karadağ et al. (2023) studied 7 different YOLOv7 (Wang et al., 2023) models on mobile devices using iOS operating system. Model parameter space ranges from 6 million up to 154 million, fps is between 286 and 44, and AP (Average Precision) is between 0.387 and 0.568 for the smallest (YOLOv7 Tiny) and largest (YOLOv7-D6) models respectively. They also point out that YOLOv7 can demonstrate weak performances for too dark, too illuminated, too small, or interlaced images.

Şengül et al. (2025) run computer vision experiments on military aircraft detection using YOLOv7, YOLOv8 (Ultralytics, 2023), and RT-DETR (Lv et al., 2024) models. They used a dataset consisting of 19514 images of 43 different military aircraft classes. RT-DETR proved effective against YOLOv7 and YOLOv8 achieving a 0.879 mAP versus the 0.829 mAP and 0.877 mAP scores of YOLOv7 and YOLOv8 respectively.

Another transformer-based study is by Karakuş et al. (2025) to segment the handwritten text lines using RT-DETR model on a Turkish handwritten dataset comprising 1610 written forms. They compared DETR model with YOLO and proved superiority by 0.925 average IoU versus 0.762 average IoU of YOLO model.

Dayıoğlu et al. (2025) performed experiments with YOLO11 (Jocher and Qiu, 2024) models (nano to xlarge) to detect Printed Circuit Board defects and demonstrated that YOLO11l model outperformed the other YOLO11 models by 0.551 mAP@0.50-0.95 and YOLO11n model reached 166 frames per second speed proving YOLO11n as a strong candidate for real-time processing of PCB quality control systems.

2. Materials and Methods

This section summarizes the datasets and object detection models used in this paper. We give a brief definition of these methods and explain their internal structures and functioning.

2.1. Open Images Dataset

Open Images Dataset Version 7 (Kuznetsova et al., 2020) is a very large-scale image dataset designed for variety of tasks including classification, detection, instance segmentation, visual relationship, local narrative, point-level labeling, multi-modal image descriptions. It contains over 9 million images with bounding boxes, local narrations, segmentation masks, and visual relationships. Open Images has 600 object classes and 16 million bounding boxes for approximately 1.9 million image samples. Bounding boxes were drawn mostly by professional human annotators which provides higher accuracy. The images are usually complicated and incorporates over 8 objects per image in average. It includes 1466 visual relationship such as two men shake hands, man is jumping, and table is wooden from 3.3 million annotations, 2.8 million object instances from 350 different object classes, 675,000 local narrations of multimodal descriptions of with text, voice, and mouse traces, 66.4 million point-level labels for 5827 classes over 1.4 million images. The total number of annotations is 61.4 million image-level labels over more than 20,000 object classes. Open Images has a very high average number of objects in images (average 8.4) and object area compared to the other image detection datasets. The dataset includes validation set, training set as well as test set. Training set has more than 9 million images where validation set and test set comprise approximately 40,000 and 125,000 images respectively. This study is using 33,538 images of 5 classes (airplane, car, cat, dog, person) from the training set of Open Images dataset. These are the most used classes in Open Images dataset and also, they are common with LaSOT and COCO datasets.

2.2. LaSOT Dataset

LaSOT (Large-scale Single Object Tracking) (Fan et al., 2019) is the second image dataset which is actually designed for long-term object tracking; however, it can also be used for image detection due to its extensive and precise bounding box ground truths. LaSOT contains nearly 4 million images in 1550 sequences over 85 object classes. Ground truth bounding box annotations were done manually by human annotators with high precision. A significant property of LaSOT is the long durations of

image sequences with an average of 83 seconds making it suitable for long-term tracking. LaSOT also provides linguistic annotations as well as visual annotations. In this study, 33,790 images from LaSOT dataset are used and these images are 13,129 Airplane (airplane-1, airplane-2, airplane-3), 7643 Boat (boat-1, boat-2), 9868 Car (car-1, car-2, car-3), and 3150 Train (train-1, train-2) class images.

2.3. YOLO12

YOLO12 (Tian et al., 2025) is a groundbreaking development compared to the previous YOLO models. YOLO12 incorporates an attention-based architecture which is different than the custom CNN models of previous YOLO models yet still preserving the speed for real-time processing. An area attention model is introduced by YOLO12 which splits the feature data into different regions (usually 16) to obtain a large receptive field and dodge high computational costs. YOLO12 also introduces R-ELAN (Residual Efficient Layer Aggregation Networks) which provides residual connections through scaling and bottleneck-like structure by feature aggregation. The optimized attention mechanism leverages FlashAttention to minimize problems of accessing memory, to reduce the depth of stacked blocks, to remove positional encoding, and to inject a 7x7 convolution inside the attention block to encode the location information. These features enable YOLO12 to support different diverse platforms while maintaining accuracy with high speed and fewer parameters than ever.

2.4. RT-DETR

RT-DETR (Real-Time DEtection TRansformer) was developed by Baidu and is a cutting-edge end-to-end object detection model providing high accuracy with reasonable speed for real time applications. It eliminates the long-suffering NMS (Non-Maximum Suppression) framework by employing a vision transformer with a convolutional backbone and a powerful multi-purpose encoder which can optimally process feature maps in multiple scales using cross- and intra-scale fusion. One of the main advantages of RT-DETR is its flexibility to different levels of inference speeds with its various decoding layers. RT-DETR consists of a powerful Hybrid Encoder with two main modules: AIFI (Attention-based IntraScale Feature Interaction) and CCFM (Cross-Scale Context Fusion Module). AIFI specializes on the intra-scale features and combines the advantages of convolution and transformer architectures to enhance the efficiency of detection using self-attention with localization and semantic evaluation. On the other hand, the next module CCFM focuses on the cross-scale features to spur the detection of objects with different sizes. CCFM is a replacement for feature pyramid networks and path aggregated networks to exploit the cross-scale feature fusion relations particularly for small and medium-sized objects. Unlike its predecessors' sequential strategy, CCFM uses a parallel strategy and concatenates the features to reduce the computational costs. Its channel-oriented attention and depth-based discrete convolutional structure help reduce the parameter space significantly while preserving the accuracy of the system.

2.5. RF-DETR

RF-DETR (Roboflow DEtection TRansformer) (Robinson, 2025) is a powerful real-time transformer-based model for object detection introduced by Roboflow team. The main goal of RF-DETR is to close the gap between the traditional fast but less accurate CNN-based models and slow-but highly accurate transformer-based models. RF-DETR is the first model to cross over the 60 mAP barrier in COCO dataset. RF-DETR tries to capture the relationships among the distinct parts of the image using the power of transformers to obtain the global theme of the image. It is an improved version of DE-DETR (Deformable-DETR) and LW-DETR (Lightweight DETR). Similar to the original DETR models, RF-DETR tries to remove the necessity for hand-engineered features, anchor-box dependency, and Non-Maximum Suppression post-processing simplifying the detection pipeline and making it more suitable and compatible to integrate into other systems.

2.6. Mask R-CNN

The Mask R-CNN method is technically an enhancement of the previous Faster R-CNN model to extend the bounding box calculation to obtain the exact shape and location of the objects by incorporating a parallel mask branch. The main purpose of Mask R-CNN is segmentation tasks; however, it can also be used for bounding box calculation and object detection. Mask R-CNN is built upon the Faster R-CNN structure which uses a method called RPN (Region Proposal Network) to implement attention and adds a second stage to process the ROIs from the RPN. Mask R-CNN uses RoIAlign algorithm instead of RoIPool of Faster R-CNN model. RoIAlign avoids the quantization by using bilinear interpolation and accurately aligns the input with the extracted features. Mask R-CNN leverages 3 parallel workflows including classification, bounding box detection, and masking branches. The key point in mask branch is the prediction of mask without causing any competition between the classes. As a summary ResNet backbone of Mask R-CNN extracts the visual features, RPN proposes regions for possible objects, RoIAlign aligns the features for the proposed regions, and finally Heads of Mask R-CNN classify, refine, and segment all proposed regions.

3. Results

This section lays down the results of object detection experiments using COCO-pretrained YOLO12X, Mask R-CNN Resnet50_fpn, RT-DETR-X, and RF-DETR-Large models on multi-object Open Images and single-object LaSOT datasets. All models are pretrained on COCO dataset. We should bear in mind that COCO is an 80-class object dataset whereas Open Images contains 600 object classes. The total number of images is 67,328 including 33,790 samples from LaSOT and 33,538 samples from

Open Images dataset. In the LaSOT dataset, only airplane, boat, car, and train classes are used whereas airplane, car, cat, dog, and person classes are selected from the Open Images dataset. It should be emphasized that Open Images dataset is a very unbalanced dataset in terms of class distributions. Moreover, some of the classes of Open Images are not available in the COCO dataset. In the LaSOT experiments, all classes are filtered out except the target class and final detection is based on the most overlapping prediction box with the ground truth box. Although LaSOT is a single object dataset, in many frames, there are many other classes as well as many instances of same class. For the Open Images dataset, we filtered out all classes except the airplane, boat, car, cat, and dog and used the most overlapping prediction boxes with the available ground truth boxes.

Experiments are conducted on a system with an Intel i9 14900KF CPU, RTX 4060Ti 16 GB GPU, 32 GB RAM, 1 TB SSD, 2 TB HDD, and Torch 2.2.0. The IoU and confidence score thresholds are fixed at 0.5 and 0.7 respectively for all models in all experiments. Model size and speed comparisons are given in Table 1. As can be seen from Table 1, YOLO12X is the fastest and smallest model with 49.78 fps and 113 MB size followed by RT-DETR-X with 28.64 fps and 129 MB size. RF-DETR-Large is the largest model by far with a size of 1.46 GB and the second slowest following Mask R-CNN. The fps values are GPU-based.

Table 1. Size (MegaBytes) and fps (frame per second) of the models used in the experiments on LaSOT dataset

	Size (MB)	fps
YOLO12X	113	49.78
Mask R-CNN	178	12.38
RT-DETR-X	129	28.64
RF-DETR-Large	1460	16.20

Table 2. Number of frames (%) without prediction on LaSOT dataset

	Frames w/o prediction on LaSOT dataset
YOLO12X	0.2154
Mask R-CNN	0.0490
RT-DETR-X	0.1019
RF-DETR-Large	0.1503

Detectors can generate different behaviors in different classes and cannot produce a prediction for the target frame even though the object is very clearly visible on the image. In Table 2, frames without predictions are presented for all models. As can be seen from Table 2, Mask R-CNN posts the lowest miss rate followed by RT-DETR-X with 0.0490 and 0.1019 miss rates respectively. YOLO12X has the highest miss rate with 0.2154. In the LaSOT dataset, YOLO12X is missing 0.5499 of the frames from the boat class without generating any prediction and bounding box as depicted in Figure 1. YOLO12X is also missing 0.2733 of the frames from the train class without any predictions as depicted in Figure 2. RF-DETR-Large is missing 0.2867 of the frames from the airplane class which is actually a helicopter as illustrated in Figure 3. RT-DETR-X also misses 0.1898 of frames and Mask R-CNN misses 0.2968 of frames from the train class. In train class of LaSOT dataset, the ground truth box contains only the locomotive; however, detectors include the full train in their bounding-boxes as in Figure 4.

Figure 1. YOLO12X is missing too many frames without prediction from boat classes of LaSOT dataset.

Figure 2. Train class of LaSOT dataset is another YOLO12X failure example.

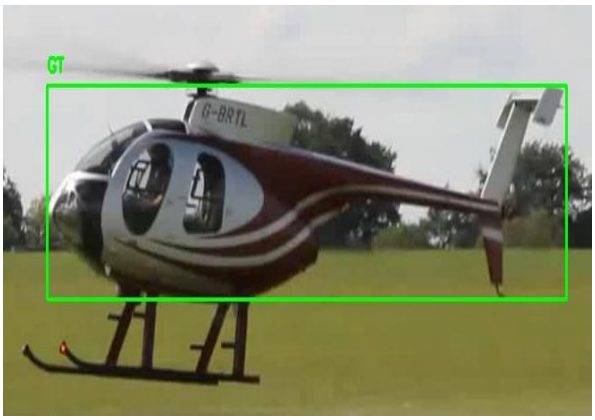

Figure 3. RF-DETR-Large is missing too many frames without producing any prediction in airplane-3 class of LaSOT dataset.

Interlacing objects inside a bounding box is a major problem for Mask R-CNN model as depicted in Figure 5. YOLO12X, RT-DETR-X, and RF-DETR-Large rarely incur interlacing problem. In the following figures, green box denotes the ground truth (GT) and red box denotes the prediction bounding box.

Open Images is a very difficult dataset for COCO-pretrained models. COCO includes 80 classes versus the 600 classes of Open Images. Some sample predictions from Open Images dataset using RT-DETR-X model are illustrated in Figure 6. Evaluation results of the experiments are tabulated in Table 3 for Open Images and in Table 4 for LaSOT datasets. It should be noted that finding a reliable evaluation metric is a very difficult choice and extreme care must be taken even including visual expert inspection of the output predictions. In this study, output predictions are also inspected visually for a better reasoning. Bar graph visualization is also presented in Figure 7 and Figure 8 for Open Images and LaSOT datasets respectively. In Open Images dataset, IoU scores are very close for YOLO12X, RT-DETR-X, and RF-DETR-Large models. RT-DETR-X scores the best mAP@0.5 and F1-score with 0.6775 and 0.5904, YOLO12X has the highest IoU and precision with 0.8636 and 0.7157 respectively, but Mask R-CNN holds the highest confidence score of 0.9516 and recall of 0.6508.

In LaSOT dataset, RT-DETR-X obtains the best IoU, F1-score and mAP@0.5, Mask R-CNN retains the best Recall and Confidence Score. As stated, defining a winner by metrics is quite difficult, however the evaluation metrics and visual inspections of the output predictions denote that RT-DETR-X is a bit ahead of the others. RT-DETR-X rarely experiences interlacing objects inside one another like Mask R-CNN and it also does not miss frames without generating an output like YOLO12X and RF-DETR-Large for long periods. RF-DETR-Large is not generating any best score in any metric on both datasets.

Figure 4. In train classes of LaSOT dataset, the ground truth encompasses only locomotive, but detectors include the whole train. Green box is the ground truth and red box is the model prediction box.

Table 3. Experimental results of multi-object detection on Open Images dataset

	IoU	Confidence	Precision	Recall	F1-Score	mAP@0.5
YOLO12X	0.8636	0.8619	0.7157	0.4576	0.5582	0.6130
Mask R-CNN	0.7920	0.9516	0.4571	0.6508	0.5370	0.3239
RT-DETR-X	0.8469	0.8897	0.6685	0.5286	0.5904	0.6775
RF-DETR-Large	0.8587	0.8848	0.6938	0.5070	0.5858	0.4235

Table 4. Experimental results of single-object detection on LaSOT dataset

	IoU	Confidence	Precision	Recall	F1-Score	mAP@0.5
YOLO12X	0.8572	0.8677	0.6309	0.7045	0.6657	0.5357
Mask R-CNN	0.8151	0.9730	0.4005	0.8622	0.5469	0.6031
RT-DETR-X	0.8804	0.9067	0.6075	0.8389	0.7047	0.6597
RF-DETR-Large	0.8582	0.9016	0.5627	0.7759	0.6523	0.6015

Figure 5. Mask R-CNN interlaces objects inside one another as in a) airplane-1, b) airplane-2 classes of LaSOT dataset.

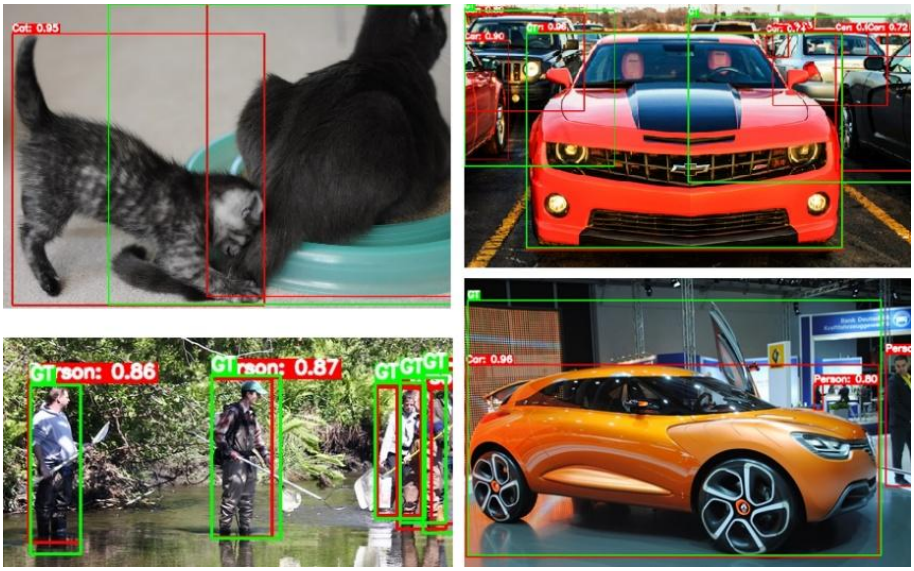

Figure 6. Sample RT-DETR-X model predictions from Open Images dataset.

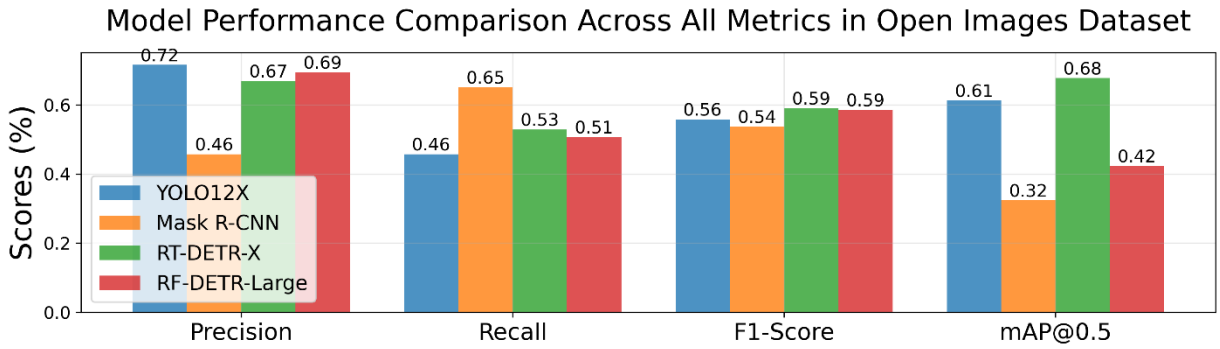

Figure 7. Complete summary of metrics of experimental results on the Open Images dataset for all models.

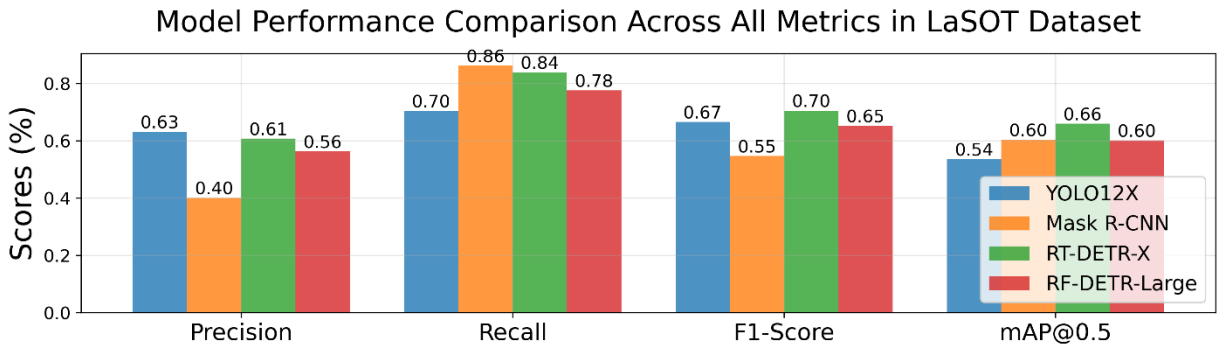

Figure 8. Complete summary of metrics of experimental results on the LaSOT dataset for all models.

4. Discussions

This manuscript presents the performance of the latest object detection models, namely, YOLO12X, Mask R-CNN with Resnet50_fpn, RT-DETR-X, and RF-DETR-Large models on multi-object Open Images and single-object LaSOT datasets. All models used in this paper are COCO-pretrained and the largest model that is currently provided by the developers. Although, all models perform very competitive, RT-DETR-X exhibits stronger performance in both datasets with respect to the mAP@0.5 and F1-scores. One of the interesting findings of the study is that some detectors like YOLO12X and RF-DETR-Large can miss object classes without outputting any prediction for long periods even if the object is very clear on the image. Another point arising from the paper is the interlacing problem of Mask R-CNN which is rarely a problem of others. It should also be noted that all models are able to generate very accurate bounding boxes and class predictions but YOLO12X and RF-DETR-Large are missing frames for long period of times and Mask R-CNN demonstrates a very severe interlacing problem. Mask R-CNN uses NMS threshold with a default value of 0.5. This can be one of the problems of interlacing. The use of 0.7 confidence threshold may need

to be increased to avoid interlacing. Mask R-CNN uses a two-stage detection algorithm and produces too many bounding boxes, therefore causing interlacing problems. RT-DETR-X uses Hybrid Encoder Attention (Efficient Multi Scale Attention) which is highly optimized compared to the previous attention models. This new attention model helps reduce the computational cost very significantly and paves the way for real-time fast and accurate object detection. RT-DETR-X also leverages Query-based Cross Attention where queries use attention segments to refine themselves. Attention heads for classification and box generation is used to decouple task and Multi-Scale Fusion with Attention is used to precisely detect large and small objects. Using these advanced attention mechanisms, RT-DETR-X eliminates the need for post-NMS. The last finding of the study is the undisputable speed advantage of YOLO12X model. In the future studies, object detection can be implemented with a model trained on a larger dataset instead of using COCO pretrained models to achieve better outcomes.

Author Contributions

The percentages of the author' contributions are presented below. The author reviewed and approved the final version of the manuscript.

	C.P.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	100

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declares that there is no conflict of interest in this study.

Ethical Approval Statement

Since no studies were conducted on animals and humans in this study, ethics committee approval was not obtained.

Acknowledgements

There is no financial support for this article.

References

Atik, M. E., Duran, Z., & Özgünlük, R. (2022). Comparison of YOLO versions for object detection from aerial images. *International Journal of Environment and Geoinformatics*, *9*(2), 87–93. https://doi.org/10.30897/ijegeo.1010741

Bakır, H., & Bakır, R. (2023). Evaluating the robustness of YOLO object detection algorithm in terms of detecting objects in noisy environment. *Journal of Scientific Reports-A*, (54), 1–25. https://doi.org/10.59313/jsr-a.1257361

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, W., Luo, J., Zhang, F., & Tian, Z. (2024). A review of object detection: Datasets, performance evaluation, architecture, applications and current trends. *Multimedia Tools and Applications*, *83*(24), 65603–65661. https://doi.org/10.1007/s11042-023-17949-4

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, (pp. 886–893). IEEE. https://doi.org/10.1109/CVPR.2005.177

Dayıoğlu, M., Eyüboğlu, A. K., & Ünal, R. (2025). Performance analysis of YOLO11 models in PCB defect detection tasks. *Kuzey Ege Teknik Bilimler ve Teknoloji Dergisi*, *2*(1), 33–50.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale [Paper presentation]. *International Conference on Learning Representations (ICLR), Virtual.*

Ereken, Ö. F., & Tarhan, Ç. (2025). Modeling objects with artificial intelligence based image processing techniques: Object detection with Mask R-CNN. *Academic Platform Journal of Engineering and Smart Systems*, *13*(1), 17–21. https://doi.org/10.21541/apjess.1542885

Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., & Ling, H. (2019). LaSOT: A high-quality benchmark for large-scale single object tracking. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 5374–5383). IEEE. https://doi.org/10.1109/CVPR.2019.00552

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (pp. 1440–1448). IEEE. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 580–587). IEEE. https://doi.org/10.1109/CVPR.2014.81

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (pp. 2961–2969). IEEE. https://doi.org/10.1109/ICCV.2017.322

Jocher, G. (2020). *Ultralytics YOLOv5* (Version 7.0) [Computer software]. GitHub. https://github.com/ultralytics/yolov5

Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11* (Version 11.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

Karadağ, B., & Arı, A. (2023). Akıllı mobil cihazlarda YOLOv7 modeli ile nesne tespiti. *Politeknik Dergisi*, *26*(3), 1207–1214. https://doi.org/10.2339/politeknik.1296541

Karakuş, O. F., Gülcü, A., & Karaca, A. C. (2025). Adapting vision

transformer-based object detection model for handwritten text line segmentation task. *Journal of Innovative Science and Engineering*, *9*(1), 28–38. https://doi.org/10.38088/jise.1471047

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2020). The Open Images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *Int J Comp Vis*, *128*(7), 1956–1981. https://doi.org/10.1007/s11263-020-01316-z

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

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 (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-0_2

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comp Vis*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., & Liu, Y. (2024). RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. arXiv. https://arxiv.org/abs/2407.17140

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 6517–6525). IEEE. https://doi.org/10.1109/CVPR.2017.690

Redmon, J., & Farhadi, A. (2018). *Yolov3: An incremental improvement*. arXiv. https://doi.org/10.48550/arXiv.1804.02767

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 779–788). IEEE. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*, 91–99.

Robinson, I., Robicheaux, P., & Popov, M. (2025). *RF-DETR* [Computer software]. GitHub. https://github.com/roboflow/rf-detr

Sun, Y., Sun, Z., & Chen, W. (2024). The evolution of object detection methods. *Engineering Applications of Artificial Intelligence*, *133*, 108458. https://doi.org/10.1016/j.engappai.2024.108458

Şengül, F., & Adem, K. (2025). Detection of military aircraft using YOLO and transformer-based object detection models in complex environments. *Bilişim Teknolojileri Dergisi*, *18*(1), 85–97. https://doi.org/10.17671/gazibtd.1549034

Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. arXiv. https://doi.org/10.48550/arXiv.2502.12524

Ultralytics. (2023). *Ultralytics YOLOv8* [Computer software]. https://github.com/ultralytics/ultralytics

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, *1*, 511–518. IEEE. https://doi.org/10.1109/CVPR.2001.990517

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 7462–7471). IEEE. https://doi.org/10.1109/CVPR52729.2023.00721

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018). DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 3974–3983). IEEE. https://doi.org/10.1109/CVPR.2018.00418

Yuille, A. L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, *3*(1), 59–70. https://doi.org/10.1162/jocn.1991.3.1.59