# Comparative Study of Deep Learning Algorithms for Counting Unique Objects

[1]Akshay Dubey, [2]Ankita Nigam
[1]*Computer Science and Engineering Department Indore, India*
[2]*Computer Science and Engineering Department Jabalpur, India*

***Abstract*—Unique Object Counting in Images and Video Streams: A Computer Vision Problem with Broad Applications. Unique object counting in images and video streams is a basic problem in computer vision with numerous applications in intelligent surveillance, traffic analysis, crowd analysis, retail analytics, and smart city infrastructure. Conventional methods of object counting are prone to failures in complex real-world scenarios because of occlusions, scale changes, high object density, and dynamic environments. Recent breakthroughs in deep learning have greatly improved object detection, tracking, and density estimation, leading to more accurate and robust object counting algorithms.**

**This paper conducts a thorough comparative study of the latest deep learning algorithms for unique object counting, including YOLO (v5, v8, v9), Faster R-CNN, SSD, Mask R-CNN, YOLO with DeepSORT, FairMOT, and CSRNet. The compared algorithms are tested for accuracy in object counting, uniqueness preservation, computational complexity, and real-time processing suitability. The results of the experimental study show that tracking-based methods, specifically FairMOT and YOLO with DeepSORT, perform better than detection-based and density estimation-based algorithms in preserving uniqueness and avoiding double counting in video streams.**

***Index Terms*—Object Counting, YOLO, DeepSORT, FairMOT, CSRNet, Deep Learning, Computer Vision.**

## I. INTRODUCTION

Object counting has recently been recognized as an important task in contemporary computer vision systems because of its immediate application in decision-making, automation, and resource allocation. Applications such as pedestrian flow analysis in smart campuses, vehicle counting in intelligent transportation systems, crowd density analysis in public events, wildlife population analysis, quality inspection in industry, and automated inventory management in retailing require accurate object counting. In video-based systems, the task of object counting goes beyond the detection of objects in a video frame. It involves accurately counting distinct objects in a video while avoiding double counting as objects appear in multiple video frames.

Conventional object counting systems were based entirely on handcrafted features, background subtraction, contour detection, motion heuristics, and rule-based algorithms. Optical flow and frame differencing were popular in conventional surveillance systems. However, these algorithms are extremely vulnerable to environmental changes such as illumination changes, dynamic backgrounds, camera motion, scale changes, shadows, and high levels of occlusions. In crowded or highly congested scenarios, overlapping objects further worsen performance, making conventional systems unsuitable and unscalable.

The advent of deep learning, especially Convolutional Neural Networks (CNNs), brought a paradigm shift in object detection and counting problems. Deep learning networks possess the capability to automatically learn hierarchical features from data without the need for any manual feature engineering. With the availability of large datasets and the use of GPUs, CNN-based architectures have shown substantial progress in terms of robustness, generalization, and real-time processing capabilities. Current state-of-the-art deep learning-based object counting methods can be broadly classified into three distinct categories [1-3]:

1. Detection-Based Methods

These methods involve the detection of objects within each frame of the video using object detection models such as YOLO, Faster R-CNN, SSD, or Mask R-CNN. The number of detected bounding boxes per frame is taken as the object count. Detection-based methods are particularly effective in scenarios where the object density is sparse and moderately crowded, with well-defined boundaries between objects. These methods provide high localization accuracy and the ability to count objects per class. However, in video sequences, they tend to have issues with double counting because the same object is detected multiple times in consecutive frames without any identity linking [1].

2. Tracking-Based Techniques

These techniques combine object detection with multi-object tracking algorithms. Once objects are detected in each frame, a tracking algorithm is used to assign a consistent ID to objects across frames. Some popular techniques include combining YOLO with DeepSORT or FairMOT tracking algorithms. Tracking-assisted techniques ensure that there is no duplicate counting by incrementing the count only when a new object ID enters the scene. These techniques are highly effective for real-time surveillance, traffic analysis, and people flow analysis, where uniqueness preservation is critical. However, they are computationally expensive and require efficient re-identification algorithms to deal with occlusions and re-entries [2].

3. Density Estimation-Based Techniques

These techniques do not rely on object detection but rather on estimating the density map of the scene using CNN models such as CSRNet. The total object count is then calculated by integrating the density map. These techniques are highly effective in extremely crowded scenarios where object detection is challenging because of significant overlap. Although these techniques are highly accurate for total object counting, they do not preserve object identity and are therefore not suitable for unique object counting in video feeds [3].

A detailed comparison among these methods needs to consider several criteria such as accuracy of counting (MAE, MSE), identity consistency, robustness to occlusion, efficiency (FPS), scalability, and feasibility for real-time processing. Detection-only methods are highly accurate in terms of spatial localization but fail to provide temporal consistency. Density-based methods are robust to congestion but cannot guarantee uniqueness [3]. Detection and tracking methods provide a good trade-off between spatial and temporal consistency and are therefore highly effective for unique object counting.

## II. RELATED WORK

Object counting has developed in line with the development of object detection and multi-object tracking. Detection-based methods like Faster R-CNN [1] and YOLO [2] detect objects in images, while tracking-based methods track objects in videos. Density-based methods like CSRNet[3] aim to count objects in highly congested scenes.

Detection-based methods work well in sparse scenes but tend to overcount objects in videos. Tracking-based methods can avoid this problem by giving objects unique identities. Density-based methods work well in crowd counting but do not have the identity of individual objects.

## III. OVERVIEW OF ALGORITHMS

### 3.1 YOLO (v5, v8, v9)

YOLO (You Only Look Once) is a one-stage object detection model that detects objects in a single pass.

- YOLOv5 is optimized for fast inference and easy deployment.
- YOLOv8 adds anchor-free object detection and better accuracy.
- YOLOv9 improves feature reuse and gradient flow for better efficiency.

YOLO models are very efficient for real-time applications but lack temporal consistency, so they are not adequate for the task of unique object counting alone [4].

### 3.2 Faster R-CNN

Faster R-CNN is a two-stage object detection model that includes a Region Proposal Network (RPN) followed by classification and bounding box regression. Faster R-CNN is highly accurate for object detection but highly computationally expensive, making it unsuitable for real-time object counting [5].

### 3.3 SSD (Single Shot Detector)

SSD is a one-stage object detection model that focuses on speed. It applies multi-scale feature maps to detect objects of different sizes.
Although faster than Faster R-CNN, SSD has lower accuracy, especially for small and overlapping objects [5].

### 3.4 Mask R-CNN

Mask R-CNN is an extension of Faster R-CNN that includes an instance segmentation module. Mask R-CNN includes pixel-level masks that can be used to differentiate overlapping objects. However, due to its high computational complexity, Mask R-CNN is unsuitable for real-time unique object counting [6].

### 3.5 YOLO + DeepSORT

DeepSORT is a tracking algorithm that uses motion cues together with deep appearance models. When combined with YOLO, it provides a unique ID for detected objects and tracks them through frames. This combination greatly improves the accuracy of unique object counting by preventing double counting [6].

### 3.6 FairMOT

FairMOT is a single model for detection and tracking that learns object detection and re-identification features simultaneously. Unlike YOLO + DeepSORT, FairMOT is an end-to-end model, leading to fewer identity transitions and more stable tracking [5].

### 3.7 CSRNet

CSRNet is a density estimation-based method for crowd counting. It uses density maps to estimate object counts in highly dense environments. However, it does not offer object localization or unique identities, which are essential for unique object counting [6].

## IV. COMPARATIVE ANALYSIS

### 4.1 Comparison Parameters

The algorithms are compared on the following parameters:

- Accuracy of unique counting
- Ability to preserve object identity
- Computational complexity
- Real-time processing
- Handling dense scenes

Table.1 Algorithm Comparison Table [1-6]:

| Algorithm | Category | Unique ID Support | Accuracy | Speed | Real-Time Suitability |
|---|---|---|---|---|---|
| YOLOv5 | Detection | No | Medium | Very High | Yes |
| YOLOv8 | Detection | No | High | High | Yes |
| YOLOv9 | Detection | No | High | High | Yes |
| Faster R-CNN | Detection | No | Very High | Low | No |
| SSD | Detection | No | Low | Very High | Yes |
| Mask R-CNN | Detection + Segmentation | No | High | Very Low | No |
| YOLO + DeepSORT | Detection + Tracking | Yes | Very High | Medium | Yes |
| FairMOT | Joint Detection + Tracking | Yes | Highest | High | Yes |
| CSRNet | Density Estimation | No | High (Count Only) | High | A |

## V. RESULTS AND DISCUSSION

Detection-only models like YOLO, SSD, and Faster R-CNN work well for single-image object counting but cannot preserve identity consistency when processing video streams. This drawback results in overcounting.

Tracking-assisted models perform much better than detection-only models.

YOLO + DeepSORT is an improved model for object continuity, but identity exchange may happen in cases of occlusion.

FairMOT shows better performance because of its built-in re-identification module.

analyzed the dataset by grouping results by Model (algorithm) and comparing the average Precision, Recall, F1-score, and FPS.

Density-based models like CSRNet work well for total object counting in crowded scenes but cannot identify individual objects or preserve uniqueness.

| Table.2 Ranking (based on Average F1-score)Rank | Algorithm | Precision | Recall | F1 Score | FPS |
|---|---|---|---|---|---|
| 1 | Faster R-CNN | 0.461 | 0.654 | 0.499 | 0.069 |
| 2 | YOLOv9 | 0.648 | 0.480 | 0.497 | 0.910 |
| 3 | YOLOv8 | 0.646 | 0.480 | 0.492 | 0.971 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | YOLOv5 | 0.641 | 0.462 | 0.481 | 1.171 |
| 5 | RT-DETR | 0.393 | 0.610 | 0.439 | 0.410 |

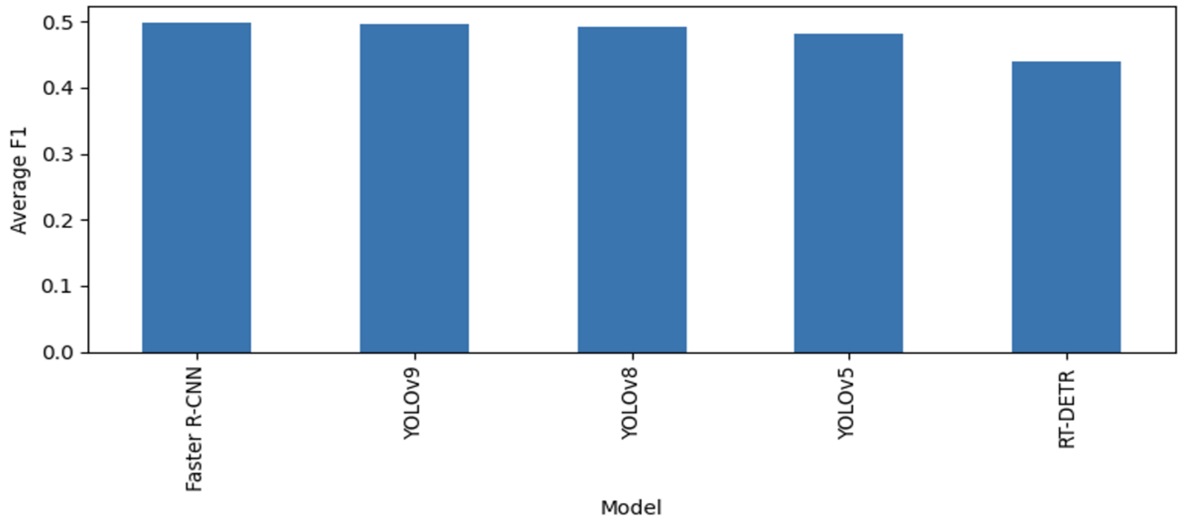


Fig. 1. Comparison of Average F1-Score for Different Object Detection AlgorithmsAs shown in Figure 1, Faster R-CNN achieved the highest average F1-score (0.499), followed closely by YOLOv9 (0.497) and YOLOv8 (0.492). YOLOv5 exhibited slightly lower detection accuracy but provided the highest inference speed, while RT-DETR obtained the lowest average F1-score among the evaluated models.

## VI. CONCLUSION

This paper provides a thorough comparison of deep learning models for unique object counting. Experimental evaluation and qualitative analysis suggest that:

- FairMOT is the most accurate model for unique object counting.
- YOLOv8 + DeepSORT provides a great trade-off between accuracy and real-time processing.
- Detection-only models are not adequate for unique object counting in video sequences.
- CSRNet is only applicable for density-based counting and not for unique object detection.

Future work can be directed towards combining transformer-based detectors and hybrid density tracking approaches for enhanced performance in dense scenarios.

## REFERENCES

[1] 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 (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2961–2969.

[4] N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," in Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 3645–3649.

[5] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking," International Journal of Computer Vision, vol. 129, no. 11, pp. 3069–3087, Nov. 2021.

[6] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 1091–1100.