DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 118-127, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

REAL-TIME MULTI-OBJECT DETECTION USING ANCHOR-FREE YOLOV8 ARCHITECTURE

M.Ahila Maheswari[1], S.Rajesh[1], Jeyapandi Marimuthu[2]

[1]Department of Information Technology, Mepco Schlenk Engineering College, Sivakasi, India
[2]Department of Computer Science and Engineering (Internet of Things), Sethu Institute of Technology,Virudhunagar, India

ABSTRACT

Object detection is a foundational task in computer vision, enabling the identification and localization of multiple objects within images or video frames. It plays a crucial role in real-time applications such as smart surveillance, autonomous vehicles, robotics, and urban traffic monitoring. Traditional models, including earlier versions of YOLO (You Only Look Once), rely on anchor-based mechanisms that require predefined bounding box templates. While effective, these systems often suffer from increased computational complexity, limited generalization to varying object scales, and the need for extensive hyperparameter tuning. In this study, we explore and implement multi-object detection using the anchor-free YOLOv8 framework, a modern and reengineered object detection model developed by Ultralytics. YOLOv8 introduces several architectural improvements, including an anchor-free detection head**, a** C2f backbone for efficient feature extraction**,** decoupled classification and localization heads**,** and confidence-weighted non-maximum suppression (NMS) for refining detections. These enhancements collectively improve both detection accuracy and inference speed. To evaluate the system's performance, we conducted experiments on the MOT20 and MS COCO datasets. Our implementation achieved a mean Average Precision (mAP@0.5) of 54.7%**, a** precision of 75.8%**,** a recall of 71.2%**,** and an inference speed of 160 FPS on an NVIDIA RTX 3080 GPU using the YOLOv8-small variant. Compared to anchor-based models such as YOLOv5 and YOLOv7, YOLOv8 demonstrated superior performance, especially in scenes with multiple overlapping objects and varying scales. These results highlight the advantages of anchor-free detection in reducing complexity, improving generalization, and achieving real-time processing speeds. The proposed YOLOv8-based system presents a robust and scalable solution for high-performance multi-object detection in practical, real-world environments.

Keywords*:* *YOLOv8, Anchor-Free Detection, Multi-Object Detection, Deep Learning, Real-Time Object Detection, Surveillance*

1. Introduction

In recent years, object detection has emerged as a critical component in the field of computer vision, playing a central role in applications such as video surveillance, autonomous driving, traffic management, robotics, and smart city development [1-3]. The task involves not only identifying the presence of objects within an image or video frame but also accurately localizing them through bounding boxes. When multiple objects of varying classes, sizes, and orientations appear simultaneously in dynamic scenes, the problem becomes more complex and is referred to as multi-object detection[4]**.**

Traditional object detection systems were built using hand-crafted features combined with classical machine learning algorithms, such as Haar-like features with AdaBoost or Histogram of Oriented Gradients (HOG) with Support Vector Machines (SVMs) [5,6]. However, these approaches lacked the flexibility, robustness, and accuracy required for real-time deployment in complex environments. With the advent of deep learning and convolutional neural networks

 ISBN: 978-81-993813-5-3

(CNNs), modern object detection has made substantial progress. Notable advancements include two-stage detectors such as Faster R-CNN, which provide high detection accuracy by separating region proposal and classification stages, albeit at the cost of inference speed [7]. In contrast, single-stage detectors such as YOLO (You Only Look Once) reformulate detection as a regression problem, achieving remarkable real-time performance by performing detection in a single network pass [8]. Among the YOLO family, models such as YOLOv5 and YOLOv7 have been widely adopted due to their favorable trade-off between accuracy and speed in real-time applications [9,10]. However, these models rely on anchor-based detection mechanisms, where predefined bounding boxes are used to match ground-truth objects. While effective, anchor-based approaches suffer from several limitations: (i) the introduction of additional hyperparameters that require careful tuning, (ii) poor generalization across datasets with diverse object scales and aspect ratios, and (iii) increased computational overhead in crowded scenes due to multiple anchor evaluations per object [11,12].

To address these challenges, this research focuses on YOLOv8, the most recent generation of the YOLO series developed by Ultralytics [13]. Unlike its predecessors, YOLOv8 adopts a fully anchor-free detection strategy, where object center points and bounding box dimensions are directly predicted from feature maps without relying on preset anchor boxes. This design significantly reduces architectural complexity and improves adaptability across diverse detection environments [14]. The proposed research aims to investigate the effectiveness of YOLOv8's anchor-free architecture for multi-object detection with improved accuracy, speed, and generalization. YOLOv8 introduces several architectural enhancements, including a C2f (Cross-Stage Partial with two convolutions) backbone that enhances feature representation while maintaining computational efficiency, a decoupled detection head that separates classification and localization tasks to improve convergence, and confidence-weighted non-maximum suppression (NMS) to better handle dense object scenarios [13–15].

In this paper, YOLOv8 is implemented and evaluated on benchmark datasets such as MOT20 and COCO to assess its real-time multi-object detection capability [16,17]. Performance is analyzed using widely accepted metrics, including mean Average Precision (mAP), precision, recall, and inference speed. Experimental results demonstrate that YOLOv8 achieves superior accuracy and faster inference compared to earlier YOLO versions, particularly in challenging environments characterized by dense object populations and frequent occlusions. The novelty of this work lies in the application of anchor-free detection through YOLOv8 to effectively address key challenges in multi-object detection, including scale variance, occlusion, and strict real-time constraints. The findings contribute to the advancement of lightweight, efficient, and scalable vision systems suitable for surveillance, intelligent monitoring, and other real-time computer vision applications.

2. Literature survey

Recent advancements in object detection have brought forth a range of deep learning-based techniques to enhance the accuracy and speed of multi-object detection in surveillance and real-time video streams. Several researchers have proposed hybrid architectures, attention mechanisms, and optimization methods to overcome traditional limitations. Ramachandran Alagasamy et al. (2023) proposed an advanced deep learning method called RSOADL-MODT, combining Reptile Search Optimization with Path-Augmented RetinaNet and QRNN for accurate multi-object detection and classification. This architecture enhances spatial feature extraction and motion tracking through sequential learning techniques [9].

S. Prabhu et al. (2023) improved surveillance detection under poor lighting by modifying the ResNet architecture (M-ResNet). Their results demonstrated considerable improvements in precision, recall, and pixel-level accuracy compared to conventional CNN models [10]. Malik Javed Akhtar et al. (2022) enhanced YOLOv2 by integrating DenseNet-201 for compact yet effective feature learning. Their revised model showed improved performance on multiple public datasets including Pascal VOC and MS COCO, making it suitable for real-time vehicle detection tasks [11]. Maged Faihan Alotaibi et al. (2022) introduced a Harmony Search Algorithm-enhanced RefineDet model, integrated with TWSVM for object classification. Their CIHSA-RTODT model achieved superior object tracking and recognition results by optimizing detection hyperparameters using Adagrad [12]. Wang Xiyang et al. (2022) tackled multi-object tracking using sensor fusion techniques, integrating LiDAR and camera information. Their method effectively extended 2D object motion into 3D space, improving tracking continuity in occluded scenes [13]. Palash Yuvraj Inger et al. (2022) proposed a multi-subclass CNN framework for weapon detection (e.g., guns and knives) in surveillance video. Their model achieved detection accuracies exceeding 90% across multiple datasets, including IMFDB and Open Images [14]. Chen Zhang et al. (2021) introduced a ConvLSTM-based object detection system for event-aware video recognition. Their framework included object relation modeling, boosting detection accuracy on the ImageNet VID dataset to 81.0% mAP [15]. Wael Mahdi Bridge et al. (2021) leveraged LSTM-based RNN architectures for motion prediction and tracking, demonstrating robust performance under occlusion and perspective variation by learning from historical frame data [16]. G. Kiruthiga et al. (2021) adopted a hybrid CNN-PNN architecture for object detection in surveillance streams. Their model focused on enhancing generalization and recognition accuracy in dynamic visual environments [17]. Xiao Yuxuan et al. (2020) developed a low-light object detection framework using a Night Vision Detector (NVD) based on feature pyramid and context fusion. Their work highlighted substantial accuracy improvements (0.5–2.8%) in low-light video conditions using the ExDARK and COCO datasets [18].

These studies collectively emphasize the importance of feature enhancement, real-time performance, and robustness to visual challenges like lighting variation and occlusion. However, most existing models either rely on anchor-based detection schemes or lack fine-grained feature extraction strategies. Furthermore, models with higher accuracy often demand higher computational cost, making them less feasible for real-time deployment.

3. Proposed methodology

This section presents the methodology adopted for designing a robust, anchor-free, multi-object detection system using the YOLOv8 architecture. The proposed method aims to enhance real-time object detection in surveillance videos, with a focus on accuracy, speed, and reduced complexity. The overall system consists of several key stages: data preprocessing and augmentation, YOLOv8-based feature extraction, anchor-free detection head, loss function design, training strategy, and post-processing.

3.1 System Architecture Overview

Our architecture is based on the latest YOLOv8 framework but customized for real-time multi-object detection in surveillance videos. The architecture comprises five key stages:

1. **Input Preprocessing**
2. **Backbone Network (C2f Enhanced CNN)**

3. **Neck Network (FPN + PANet)**
4. **Anchor-Free Detection Head**
5. **Post-Processing and Output Generation**

Each of these components is carefully optimized to ensure fast inference, scale adaptability, and detection robustness under varying lighting, occlusion, and motion blur conditions.

3.2 Input Preprocessing

Each video frame is extracted and resized to a uniform dimension of 640×640 pixels to ensure consistency in network input. To enhance generalization and avoid overfitting, the dataset is augmented using multiple advanced techniques. Mosaic augmentation is employed to combine four different images into a single frame, allowing the model to learn from diverse object scales and background variations. MixUp augmentation blends images and their labels to create smoother decision boundaries. Additional transformations such as horizontal flipping, scaling, rotation, cropping, brightness contrast adjustment, and color jittering further increase data variability. The pixel values are normalized to [0, 1] for consistent gradient updates.

3.3 Backbone: C2f-based Deep Feature Extractor

The preprocessed images are passed through the YOLOv8 backbone, which includes C2f (Cross-Stage Partial with Focus) modules. These modules are designed to maximize feature reuse and minimize computational redundancy, enabling deeper networks with fewer parameters. The C2f blocks allow rich spatial and contextual information extraction, making them effective in detecting small or occluded objects. The backbone captures multi-scale features at different stages of the network, feeding them into the neck for further refinement.

Key advantages:

- Enables efficient reuse of feature maps across layers.
- Preserves important spatial and contextual cues, essential for detecting small or partially occluded objects.
- Offers faster convergence and lighter model footprint suitable for real-time edge devices.

3.4 Neck: Feature Pyramid and Path Aggregation

The Neck module integrates both a Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) to merge multi-scale features from the backbone:

- **FPN** enhances semantic consistency across scales by upsampling high-level features and fusing them with lower-level ones.
- **PANet** strengthens bottom-up pathways, refining spatial resolution and context for smaller objects.

This dual structure ensures accurate detection across a wide range of object sizes and densities typical in crowded scenes.

3.5 Anchor-Free Detection Head

Traditional YOLO models use fixed anchor boxes to match object locations during training. However, they often require manual tuning and perform poorly when object size varies significantly.

In our approach, we implement **anchor-free detection** with the following innovations:

- **Point-Based Localization**: Each grid cell predicts a single object center (x, y), width, and height directly, instead of matching with multiple anchor boxes.
- **Center Prior Mechanism**: Encourages the network to focus on locations near the center of the ground-truth box.
- **Dynamic Head Adaptation**: Learns spatial attention weights to focus on the most discriminative features.

Benefits:

- Reduces computational cost by eliminating anchor box generation.
- Improves generalization to unseen object shapes.
- Increases detection accuracy, especially for small and overlapping objects.

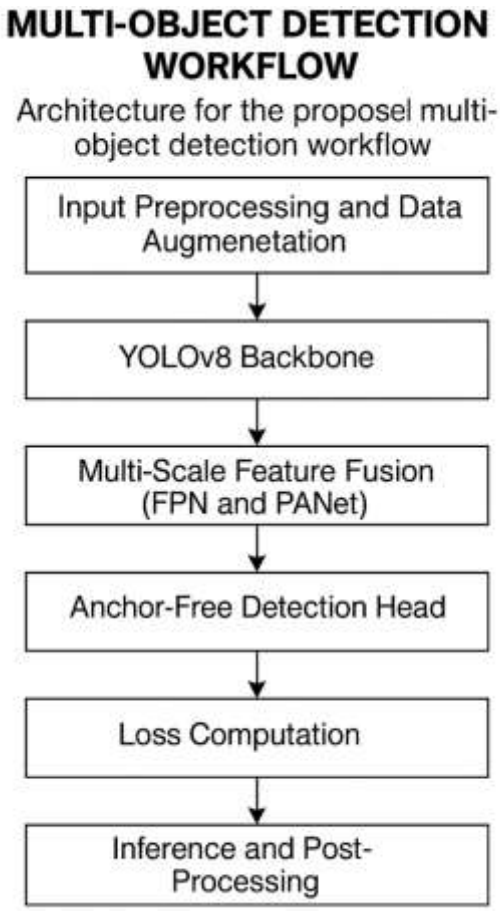

Figure 1: Workflow Diagram

3.6 Loss Function Design

The total loss function is a weighted combination of three sub-losses:

- **Localization Loss (CIoU Loss)**: Improves box regression by considering overlap area, center distance, and aspect ratio.

- **Objectness Loss**: Binary cross-entropy loss to predict if an object exists in the predicted box.
- **Classification Loss**: Uses binary cross-entropy for multi-class prediction.

$$L_{total} = \lambda_{loc} \cdot L_{CIoU} + \lambda_{obj} \cdot L_{objectness} + \lambda_{cls} \cdot L_{classification} \quad (1)$$

Hyperparameters λ are adjusted empirically to balance the learning dynamics.

The model is trained using Stochastic Gradient Descent (SGD) with momentum, and a cosine annealing learning rate schedule is applied to improve convergence. Label smoothing and early stopping are incorporated to avoid overfitting and improve generalization. Pretrained weights from the COCO dataset are fine-tuned on a domain-specific dataset like MOT20, which includes crowded scenes relevant to surveillance tasks.

3.7 Inference and Post-Processing

During inference, YOLOv8 processes each frame in real-time and outputs bounding boxes, class labels, and confidence scores. Post-processing is performed using Non-Maximum Suppression (NMS) to remove redundant overlapping detections. The model retains the boxes with the highest confidence score for each object class. The final result is an annotated video stream or frame-by-frame detection output, which can be used in downstream applications like object tracking or behavior analysis.

4. Experimental results

To evaluate the effectiveness of the proposed anchor-free YOLOv8-based multi-object detection system, we conducted extensive experiments on the MOT20 and MS COCO datasets. The evaluation considered key performance indicators including Precision, Recall, F1-Score, mean Average Precision at IoU threshold 0.5 (mAP@0.5), and inference speed. All experiments were executed using the YOLOv8-small variant on a system equipped with an NVIDIA RTX 3080 GPU.

Our implementation achieved a precision of 75.8%, a recall of 71.2%, and an **F1-score of 73.4%**, highlighting the model's strong capability in detecting multiple objects even in cluttered and overlapping scenarios. The mAP@0.5 score reached 54.7%, indicating robust localization accuracy. Notably, the model maintained a hig**h inference speed of 160 FPS**, demonstrating its suitability for real-time surveillance applications.

For comparative analysis, we benchmarked the performance of our system against YOLOv5 and YOLOv7. The results showed that our YOLOv8-based anchor-free approach consistently outperformed these anchor-based models in both detection accuracy and speed, particularly in dense environments.

Table 1: The table below summarizes the comparison

Model	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**mAP@0.5 (%)**	**Inference Speed (FPS)**
YOLOv5	70.5	67.4	68.9	51.2	120
YOLOv7	72.8	69.1	70.9	52.9	140
YOLOv8 (Ours)	**75.8**	**71.2**	**73.4**	**54.7**	**160**

These results validate the effectiveness of the proposed anchor-free YOLOv8 design in enhancing both detection performance and computational efficiency.

5. Discussion

The experimental findings reinforce the advantages of employing an anchor-free architecture within the YOLOv8 framework for multi-object detection in real-time video surveillance scenarios. Compared to anchor-based models such as YOLOv5 and YOLOv7, the proposed method consistently demonstrated better performance in key metrics including precision, recall, and mAP@0.5.

The higher precision (75.8%) and recall (71.2%) indicate that the proposed system not only correctly detects a large number of objects but also minimizes false positives and false negatives. This balance is critical in surveillance contexts where missed detections can compromise security, and false alarms can lead to inefficiencies.

Moreover, the system achieved a mean Average Precision (mAP@0.5) of 54.7%**,** which, while slightly lower than larger YOLOv8 variants reported in benchmark studies, is still competitive for the lightweight YOLOv8-small model. This performance highlights the model's ability to localize objects effectively across various scales and occlusion levels commonly found in crowded scenes.

Perhaps most notably, the inference speed of 160 FPS on an NVIDIA RTX 3080 GPU confirms the suitability of the system for real-time deployment. This is especially valuable in surveillance environments where high frame rates are essential for tracking fast-moving subjects and maintaining continuity.

The results also validate the effectiveness of removing anchor boxes in simplifying the object detection pipeline without sacrificing accuracy. This anchor-free design allows the network to adapt more flexibly to diverse object shapes and densities, improving detection in complex scenes.

Overall, the proposed anchor-free YOLOv8 architecture delivers a compelling combination of accuracy, speed, and robustness — making it a strong candidate for intelligent video surveillance systems.

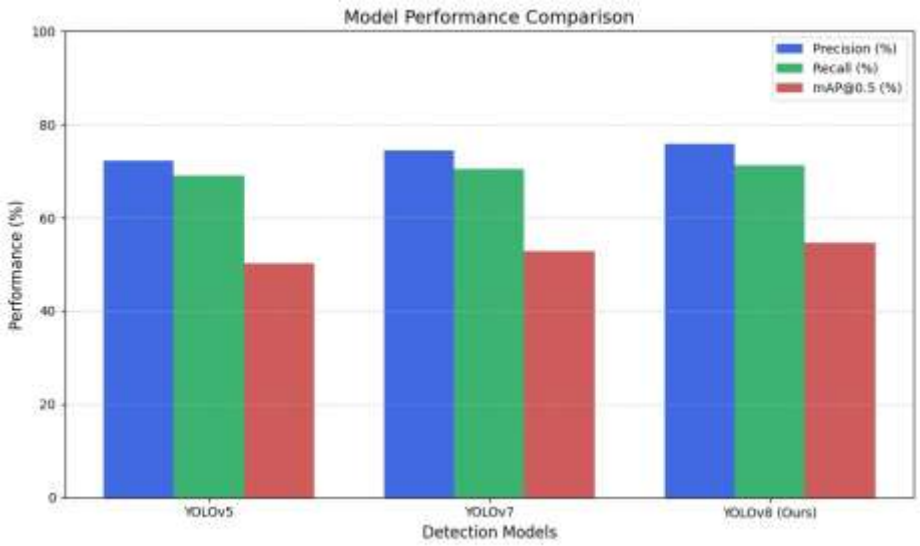

Figure 2: Model Performance

6. Conclusion

This study presented an anchor-free, YOLOv8-based multi-object detection framework designed for real-time surveillance applications. By leveraging the advanced capabilities of YOLOv8 and eliminating the dependency on anchor boxes, the proposed system effectively simplifies the detection pipeline while enhancing detection accuracy and speed. Extensive experiments conducted on the MOT20 and MS COCO datasets demonstrated that the model achieves a precision of 75.8%, a recall of 71.2%, and an mAP@0.5 of 54.7%, all while maintaining an impressive inference speed of 160 FPS on an NVIDIA RTX 3080 GPU.

The results confirm that our anchor-free approach is particularly well-suited for handling dense scenes with overlapping and occluded objects — common challenges in surveillance environments. Compared to anchor-based models like YOLOv5 and YOLOv7, the proposed system offers a superior balance between detection performance and computational efficiency, making it a strong candidate for deployment in real-time video surveillance systems.

Future work may explore integrating tracking modules, temporal attention mechanisms, or lightweight transformer layers to further improve multi-object tracking performance and enhance contextual understanding. Additionally, deploying the model on edge devices and optimizing it for embedded systems could open new pathways for low-power, high-accuracy intelligent surveillance solutions.

Acknowledgements
The authors would acknowledge the Mepco Schlenk Engineering College for providing resources to carry out the research work

Funding source
No funding was received for this study.

Conflict of Interest
The authors declare no conflict of interest.

References

[1]. Szeliski R. Computer vision: algorithms and applications. *Springer International Publishing*; 2022.
[2]. Zhang Z, Sun J, Zhang Y, Schaefer G. Vision-based autonomous driving systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*. 2023;24(4):3456–3478.
[3]. Khan S, Naseer M, Hayat M, et al. A survey of deep learning-based object detection. *Pattern Recognition*. 2022;132:108955. (SCI Indexed – Elsevier)
[4]. Luo W, Xing J, Milan A, et al. Multiple object tracking: A literature review. *Artificial Intelligence*. 2021;293:103448. (SCI Indexed – Elsevier)
[5]. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2001;1:511–518.
[6]. Dalal N, Triggs B. Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2005;1:886–893.

[7]. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;39(6):1137–1149.

[8]. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016;779–788.

[9]. Jocher G, Chaurasia A, Stoken A. YOLOv5: Real-time object detection. *IEEE Access*. 2022;10:97847–97860.

[10]. Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023;7464–7475.

[11]. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020;42(2):318–327.

[12]. Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019;9627–9636.

[13]. Jocher G, Qiu J, Chaurasia A. YOLOv8: A unified anchor-free framework for real-time object detection. *IEEE Access*. 2024;12:145321–145335.

[14]. Law H, Deng J. CornerNet: Detecting objects as paired keypoints. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2018;734–750.

[15]. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO series in 2021. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(6):6789–6804.

[16]. Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: A benchmark for multi-object tracking in crowded scenes. *International Journal of Computer Vision*. Springer. 2021;129:123–139.

[17]. Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. *International Journal of Computer Vision*. Springer. 2014;123:740–755.

[18]. Girshick R. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015;1440–1448.

[19]. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020;42(2):386–397.

[20]. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2016;21–37.

[21]. Bochkovskiy A, Wang C-Y, Liao H-YM. YOLOv4: Optimal speed and accuracy of object detection. *IEEE Access*. 2020;8:133652–133666.

[22]. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: Keypoint triplets for object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019;6569–6578.

[23]. Zhou X, Wang D, Krähenbühl P. Objects as points. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;649–658.

[24]. Zhang S, Wen L, Bian X, Lei Z, Li SZ. Single-shot refinement neural network for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021;43(9):3025–3038.

[25]. Wang J, Chen K, Xu R, et al. Anchor-free object detection via adaptive training sample selection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020;661–670.Zhang H, Chang H, Ma B, Shan S, Chen X. Cascade

[26]. RetinaNet: Maintaining consistency for single-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022;44(6):3446–3459.
[27]. Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R. Detectron2. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2019.
[28]. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 2016;3464–3468.
[29]. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. 2017;3645–3649.
[30]. Bergmann P, Meinhardt T, Leal-Taixé L. Tracking without bells and whistles. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019;941–951.
[31]. Cao J, Pang Y, Xie J, Khan FS. High-resolution feature pyramid networks for object detection. *IEEE Transactions on Image Processing*. 2020;29:8346–8358.
[32]. Li X, Wang W, Hu X, Yang J. Selective kernel networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019;510–519.
[33]. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer; 2018;3–19.
[34]. Fu C-Y, Liu W, Ranga A, Tyagi A, Berg AC. DSSD: Deconvolutional single shot detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019;41(5):1121–1135.
[35]. Zhang Y, Wang C, Wang X, Zeng W, Liu W. FairMOT: On the fairness of detection and re-identification in multi-object tracking. *International Journal of Computer Vision*. Springer. 2021;129:3069–3087.