SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I10P106

Volume 11 Issue 10, 53-60, October 2024
© 2024 Seventh Sense Research Group®

Original Article

Multi-Camera Person Tracking: Integrating YOLOv8 with ByteTrack

Nirali Anand Pandya[1], Narendrasinh C. Chauhan[2]

[1]*Gujarat Technological University, Gujarat, India.*
[2]*Information Technology Department, A. D. Patel Institute of Technology, Gujarat, India.*

[1]*Corresponding Author : thakkarniralis@gmail.com*

Received: 04 August 2024 Revised: 05 September 2024 Accepted: 05 October 2024 Published: 30 October 2024

Abstract *- Accurate and efficient person tracking in complex, multi-camera environments remains challenging. This paper proposes a novel approach that integrates the strengths of YOLOv8, an advanced model for object detection, with ByteTrack, an advanced multi-object tracking algorithm. The proposed framework is evaluated on the challenging Multi-camera Pedestrians Video Dataset to assess its performance in complex real-world scenarios. Experimental results demonstrate the effectiveness of the proposed method in accurately tracking pedestrians across multiple cameras, outperforming existing state-of-the-art techniques. Integrating YOLOv8 and ByteTrack enables robust pedestrian detection and tracking, even in challenging conditions such as occlusions, varying illumination, and camera perspectives. The proposed approach holds significant potential for intelligent surveillance systems, crowd analysis, and autonomous vehicle applications.*

Keywords *- Multi-camera Person tracking, YOLOv8, Bytetrack, Object detection, Deep Neural Network.*

1. Introduction

Accurately and efficiently tracking individuals across multiple cameras in complex, real-world environments remains a formidable challenge in computer vision. Despite significant advancements in recent years, robust person tracking continues to be hindered by a multitude of factors, including occlusions and poor lighting conditions, camera perspectives, and changes in pedestrian appearance [11]. These challenges are further exacerbated in scenarios involving dynamic backgrounds and low-resolution video footage. However, the ability to reliably identify and track individuals across multiple cameras is indispensable for a wide range of applications, such as surveillance systems, crowd analysis, human-computer interaction, and autonomous vehicles [13, 14].

Despite recent advancements in multi-camera person tracking, several challenges remain unresolved. Existing solutions often struggle with handling occlusions, where the visibility of individuals is compromised due to overlaps or objects obstructing the camera's view. These occlusions frequently result in tracking errors. Another persistent issue is the impact of varying lighting conditions. Current tracking systems perform inadequately under extreme lighting variations, such as transitioning from indoor to outdoor environments or operating under low-light conditions. This paper introduces a new method that combines the advantages of cutting-edge object detection techniques with multi-object tracking algorithms to overcome these challenges. By combining the robust detection capabilities of YOLOv8 [22] with the efficient association and re-identification techniques of ByteTrack [6], we aim to develop a comprehensive framework capable of handling the intricacies of multi-camera person tracking. Our proposed method seeks to improve upon existing approaches by addressing key limitations such as occlusion handling and identity preservation.

To evaluate the performance of our proposed framework, we conduct extensive experiments on the challenging Multi-camera Pedestrians Video Dataset [10]. This dataset is characterized by its diverse scenarios, including varying camera viewpoints, lighting conditions, and pedestrian densities, making it an ideal benchmark for assessing the robustness and generalizability of our approach. By comparing our method to existing state-of-the-art techniques, we aim to demonstrate its superior performance in terms of accuracy, efficiency, and robustness.

The structure of the paper is as follows: Section 2 offers an in-depth review of related work on multi-camera person tracking, emphasizing the advantages and drawbacks of current methods. Section 3 delves into the proposed methodology, detailing the integration of YOLOv8 and ByteTrack and the specific techniques employed to address the challenges of multi-camera tracking. Section 4 presents the experimental results, including a quantitative evaluation of our method's performance on the multi-camera pedestrians video dataset and a comparative analysis with other state-of-the-art techniques. Finally, Section 5 summarizes the key contributions of this work, discusses the limitations of our approach, and outlines potential avenues for future research.

2. Related Work

Multi-camera person tracking is crucial in computer vision, particularly for surveillance, security, and behavioral analysis. The primary objective is to maintain consistent identities of people across multiple camera views, which is challenging due to variations in viewpoints, lighting conditions, occlusions, and other environmental factors. Early works in multi-camera person tracking relied heavily on handcrafted features and classical methods such as Kalman filters and particle filters for tracking across different cameras. In one of the pioneering studies, Khan and Shah (2003) proposed a probabilistic model to estimate the 3D trajectory of a person across multiple cameras by combining 2D trajectories from individual cameras [1]. Feature-based methods involve extracting robust features like color histograms, edge descriptors, or SIFT/SURF features that can be matched across camera views. Javed et al. (2005) developed a multi-camera tracking system that uses color histograms and a homography-based appearance model to track individuals across non-overlapping camera views. These methods, however, often struggle with large variations in lighting and viewpoint [2].

The advent of deep learning has significantly advanced multi-camera person tracking. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been employed to learn discriminative appearance features and temporal dependencies. Zhang et al. (2017) introduced a deep learning-based approach that utilizes a Siamese CNN for person re-identification, which is a critical sub-task in multi-camera tracking [3]. You Only Look Once (YOLO) models have gained prominence in object detection and tracking tasks due to their real-time performance. The YOLOv8 model, the latest in the YOLO series, offers improved accuracy and speed compared to its predecessors like YOLOv3 [15], YOLOv4 [5], YOLOv5 [16], YOLOv6 [17], YOLOv7 [18]. Redmon et al. (2016) originally introduced YOLO for single-camera object detection, but its extension to multi-camera systems has shown promising results. Integrating YOLO with tracking algorithms like Simple Online and Real-time Tracking (SORT) or ByteTrack has been an emerging trend [4, 6].

ByteTrack, a recent advancement in multi-object tracking, is known for its simplicity and effectiveness. It improves upon the basic SORT [8] and DeepSORT [7] by incorporating low-confidence detections, often discarded in other algorithms, thereby enhancing tracking robustness. In the context of multi-camera systems, ByteTrack has been combined with various object detection models to maintain consistent identities across different views.

A recent study demonstrated that ByteTrack, when used with YOLOv8 [22], can significantly improve tracking accuracy in complex environments. Multi-camera person tracking has witnessed significant progress in integrating deep learning techniques. While challenges persist, ongoing research is pushing the boundaries of this field.

3. Proposed Methodology

The multi-camera person tracking system integrates a YOLOv8 object detector with a tracking pipeline using ByteTrack. Figure 1 outlines a comprehensive multi-camera tracking system leveraging the power of YOLOv8 for object detection and a sophisticated tracking pipeline (potentially ByteTrack) for maintaining identity across different camera views. Using a Kalman filter for track prediction, score-based filtering, and a two-step association process ensures the system can handle challenging scenarios such as occlusions, varying lighting conditions, and camera perspectives.

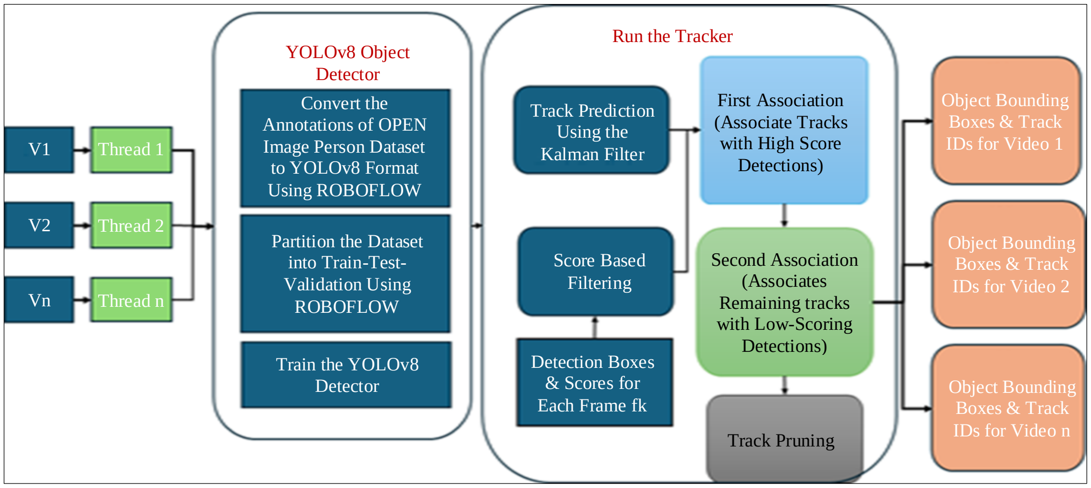

V1
Thread 1
V2
Thread 2
Vn
Thread n
YOLOv8 Object Detector
Convert the Annotations of OPEN Image Person Dataset to YOLOv8 Format Using ROBOFLOW
Partition the Dataset into Train-Test-Validation Using ROBOFLOW
Train the YOLOv8 Detector
Run the Tracker
Track Prediction Using the Kalman Filter
Score Based Filtering
Detection Boxes & Scores for Each Frame fk
First Association (Associate Tracks with High Score Detections)
Second Association (Associates Remaining tracks with Low-Scoring Detections)
Track Pruning
Object Bounding Boxes & Track IDs for Video 1
Object Bounding Boxes & Track IDs for Video 2
Object Bounding Boxes & Track IDs for Video n

Fig. 1 Proposed YOLOv8 and bytetrack for multi-camera object tracking

Figure 1 starts with multiple video streams labelled V1, V2,..., and Vn, representing input from different cameras in a multi-camera setup. Each video stream is processed in parallel using separate threads (Thread 1, Thread 2,..., Thread n), allowing the system to handle multiple video feeds simultaneously. This approach is especially beneficial for managing multiple video streams, like those from several surveillance cameras, as simultaneous processing can significantly improve efficiency and performance. Figure 2 shows the algorithm steps for the Muti-camera person tracking using YOLOv8 and ByteTrack using multi-threading.

The first step involves converting the annotations of the Open Images Dataset (which contains person annotations) to a format compatible with YOLOv8 using RoboFlow [12]. The dataset is then split into training, testing, and validation sets using RoboFlow to prepare it for training the YOLOv8 model. Once the dataset is prepared, the YOLOv8 object detector is trained. YOLOv8 is a state-of-the-art real-time object detection model known for its speed and accuracy, which is particularly useful for person detection in real-time applications.

```
Algorithm: Pseudo-code of YOLOv8 and ByteTrack-based Algorithm for Multi-Camera Object Tracking
Input:
  • Object Detector (YOLOv8) with detection threshold $\boldsymbol{\theta}$
  • Video sequence $\boldsymbol{V_1, V_2, \ldots, V_n}$
  • Number of input videos $\boldsymbol{n}$
Output: Tracks $\boldsymbol{T_1, T_2, \ldots, T_n}$for each video
    tracker_run($DetV8, \theta, V$)
        Initialize Track $T = \emptyset$
        For each frame $f_i$ in the video V:
            Obtain detection boxes and scores using the YOLOv8 model:
            $D_k = DetV8(f_i)$
            Initialize two sets: $D_H = \emptyset, D_L = \emptyset$
            For each detection $d$ in $D_K$:
              If $d.score > \theta$, add $d$ to $D_H$
              Otherwise, add $d$ to $D_L$
        Update Tracks:
        For each track $t$ in $T$, predict its new position using a Kalman Filter:
        $t = \text{Kalman}Filter(t)$
        First Association:
        Associate objects in $T$ with detections in $D_H$, using Re-ID feature distances.
        Track Unmatched Detections:
        $D_{remain} = Remaining\ object\ detections\ form\ D_H$
        $T_{remain} = Remaining\ object\ detections\ form\ T$
        Second Association:
        Match the remaining tracks in $T_{remain}$ with detections in $D_L$, using IoU similarity.
        $T_{re-remain} = Remaining\ object\ detections\ form\ T_{remain}$
        Update Unmatched Tracks:
        Remove unmatched tracks from the set:T= $T/T_{re-remain}$
        Return the updated tracks $V$ .
    tracker-thread-1 → tracker_run(Det, ϴ, V1)
    tracker-thread-2 → tracker_run(Det, ϴ, V2)
```

Fig. 2 Algorithmic Steps YOLOv8 [22] and ByteTrack [6] for multi-camera object tracking

After detecting objects (persons) in each video frame, the tracker predicts the next position of each track using a Kalman filter. The Kalman filter is a common tool in tracking systems for predicting future positions based on previous states. The tracker then performs the first association step, associating the predicted tracks with the current high-confidence detections. This step is crucial for maintaining the identity of tracked persons across frames. The detections are filtered based on a scoring mechanism, likely to remove low-confidence detections. The second association step is where the tracker associates the remaining tracks with low-confidence detections. This step helps in recovering from potential tracking failures or occlusions. Tracks deemed unreliable or no longer relevant are pruned, ensuring that the system remains efficient and focuses on valid tracks. Finally, the system outputs the object bounding boxes and corresponding track IDs for each video stream (Video 1, Video 2, ..., Video n). This output is used to identify and track individuals across different camera views.

4. Results and Discussion

4.1. Datasets

Initially, images from the Open Image Dataset [9] are used to train the YOLOv8 object detector specifically for the "Person" category. A total of 1,000 images containing 4,036 annotations for the Person class are selected. These images are divided into 70% for training, 20% for validation, and 10% for testing. Roboflow is employed to convert the annotations into YOLOv8 format and to split the labels and images into training, testing, and validation sets. The image below is an example from the dataset with annotated persons.

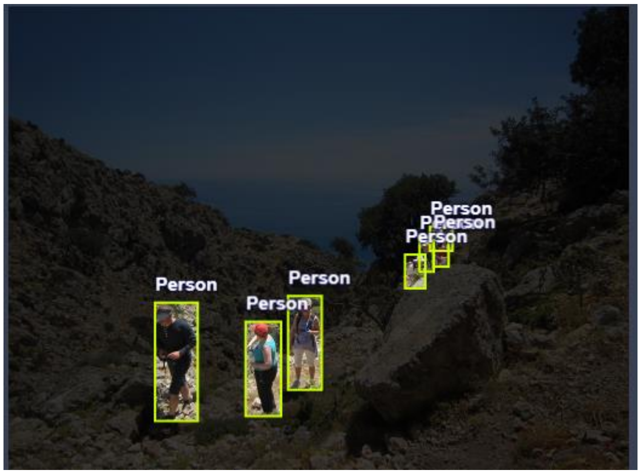


Fig. 3 Sample annotated image taken from Open Image dataset V6

We utilized sequences from the Multi-camera Pedestrian Dataset by CVLAB - EPFL [10], which includes overlapping camera views and a Real-time Multi-camera Person Dataset with non-overlapping camera views.

For the laboratory sequence of the Multi-camera Pedestrian Dataset by CVLAB - EPFL, all cameras were positioned approximately 2 meters above the ground. These sequences were recorded inside a laboratory using 4 cameras, capturing individuals as they entered the room and walked around for about 2.5 minutes. The videos were recorded at 25 Frames per Second (FPS) and encoded using the MPEG-4 codec. The passageway sequence from the same dataset was filmed in an underground passageway leading to a train station, using 4 DV cameras at 25 fps, with encoding done via the Indeo 5 codec. This sequence is challenging due to its poor lighting conditions.

The real-time multi-camera person dataset, featuring non-overlapping camera views, was captured inside a laboratory using 2 cameras. Individuals entered the room sequentially and walked around for one minute. One camera was positioned at the main entry of the premises, while the other monitored a laboratory passage with dim lighting conditions.

4.2. Performance Metrics

Performance metrics for object detection and tracking are essential for evaluating and comparing different models and algorithms. These metrics help assess the system's accuracy, efficiency, and robustness in detecting and tracking objects across video frames.

For object detection, key metrics include Precision (ratio of true positives to all detections), Recall (ratio of true positives to actual positives), IoU (overlap between predicted and ground truth bounding boxes), and mAP (mean Average Precision across classes) [21]. For tracking, metrics include MOTA (accounts for false positives and false negatives) and MOTP (measures tracking precision) [11, 19, 20].

4.3. Implementation Detail and Results

All modules were programmed in Python version 3.10.12. The deep learning models were implemented using the PyTorch framework with torch-2.2.1 and cu121 and run on a Tesla T4 GPU with 15102MiB memory. YOLOv8 and ByteTrack were implemented on Google COLAB PRO. YOLOv8 was trained on the Open Image Dataset for the "Person" category, with weights initialized from the COCO pre-trained model. The YOLOv8 model summary includes 225 layers, 3,011,238 parameters, 3,011,222 gradients, and 8.2 GFLOPs. ByteTrack utilizes a high threshold of 0.5 for the first association and a low threshold of 0.2 for the second association.

Table 1. Performance evaluation of fine-tuned YOLOv8 on Open Image dataset

Size	GFLOPs	Precision	Recall	mAP@0.5
640	8.1	0.64	0.521	0.55

Table 1 showcases the performance of the fine-tuned YOLOv8 model. The model achieves a precision of 0.64, a recall of 0.521, and a mean average precision (mAP) of 0.55 at a 0.5 IoU threshold.

Table 2. Recommended font performance evaluation ByteTrack on multi-camera Pedestrian dataset and real-time multi-camera person dataset with non-overlapping camera view

Method	Dataset	MOTA	MOTP
K-Shortest Paths [10]	EPFL Passageway Sequence [10]	68%	82%
	EPFL Laboratory Sequence [10]	70%	83%
POM [10]	EPFL Passageway Sequence [10]	68%	70%
	EPFL Laboratory Sequence [10]	72%	78%
Proposed YOLOv8 + ByteTrack	EPFL Passageway Sequence [10]	70.25%	83.3%
	EPFL Laboratory Sequence [10]	72.14%	84.6%
	Real-time Multi-camera person dataset with non-overlapping camera view	75.60%	86.8%

The plots in Figure 4 provide an in-depth analysis of YOLOv8's performance, showcasing the trade-offs between precision and recall at varying confidence levels. The mAP plot highlights the overall detection accuracy.

Figure 5 illustrates the convergence of training and validation losses for the YOLOv8 object detector and classifier after 100 epochs on the Open Image dataset. Figure 6 demonstrates the model's effectiveness in an underground passageway with challenging lighting conditions using multiple cameras. Figure 7 showcases tracking performance in a controlled indoor environment with multiple cameras. Figures 8 and 9 display the tracking results of the Multi-camera Person Dataset with Non-overlapping Camera View for the Laboratory sequence and Main entry, respectively. Table 2 presents a comparison of different methods, including the proposed YOLOv8 and ByteTrack, evaluated across various datasets. The proposed approach achieves the highest MOTA and MOTP scores in all cases, reflecting superior tracking accuracy and precision. It excels on the Real-time Multi-camera Person Dataset with a Non-overlapping Camera View, achieving a MOTA of 75.60% and a MOTP of 86.8%.

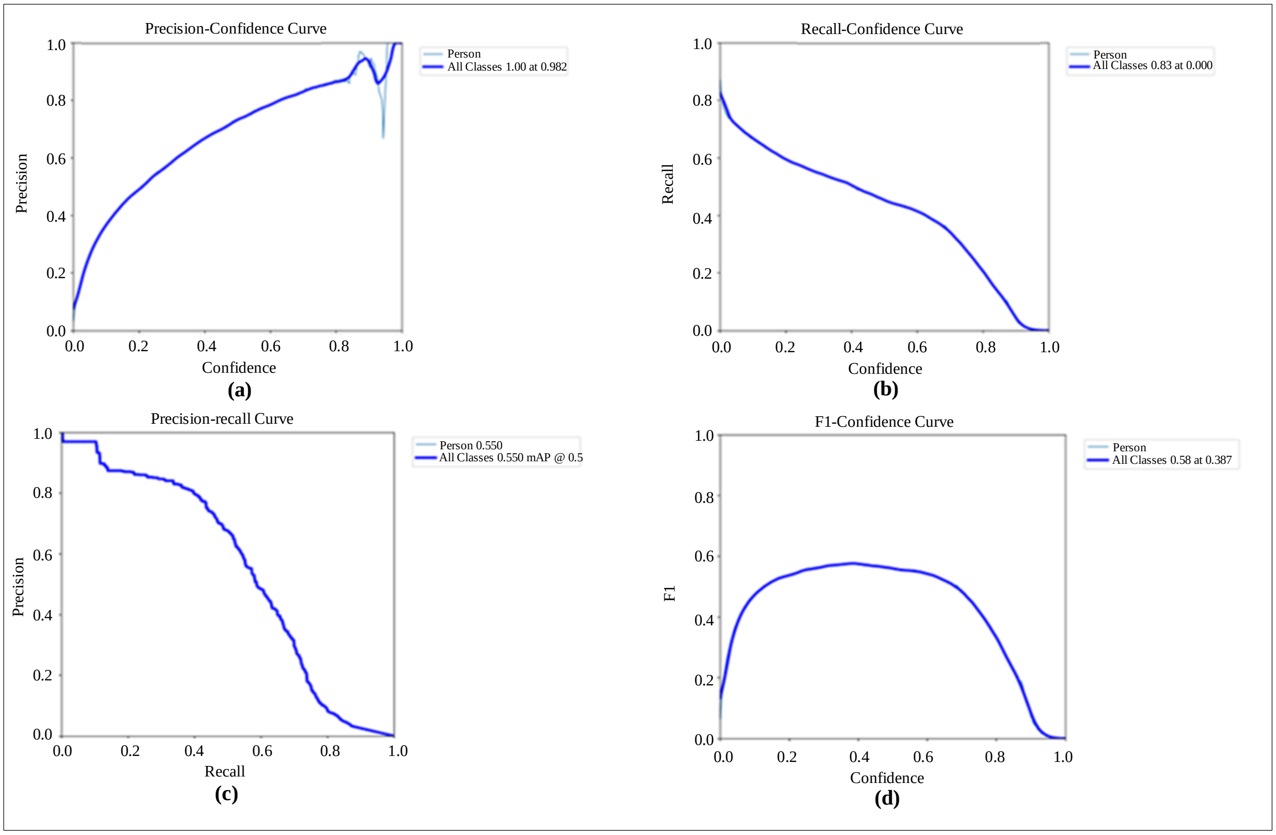


Fig. 4 presents the following: (a) precision (P) plotted against confidence (C), (b) recall plotted against confidence, (c) the mean average precision, determined by comparing detected bounding boxes with ground truth bounding boxes, and (d) the IDF1 score.

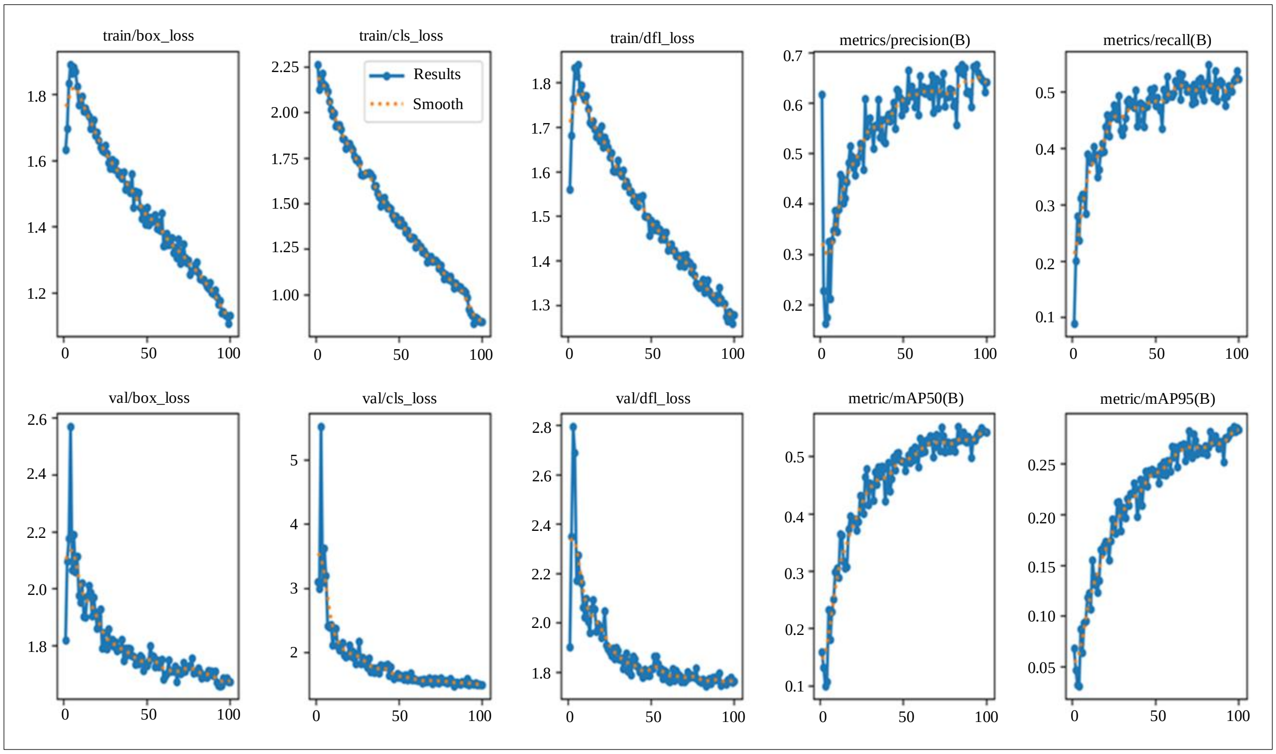
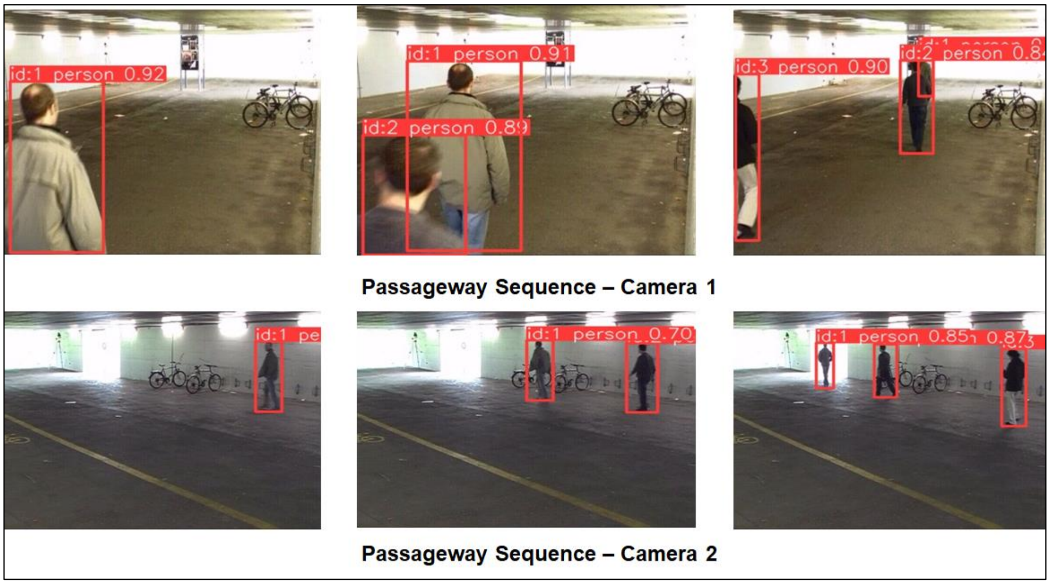


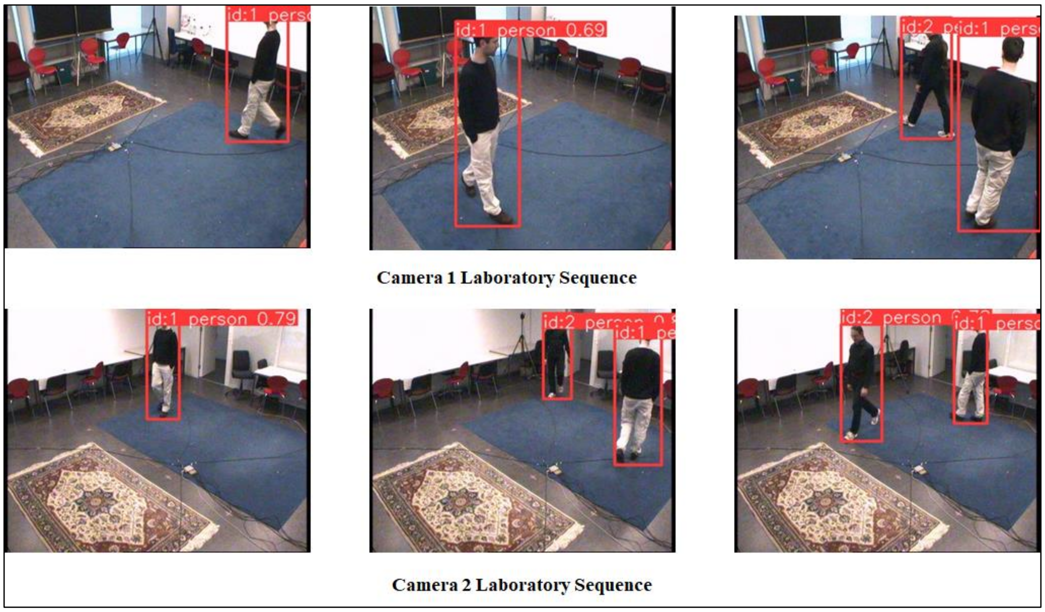
Fig. 5 The convergence of both training and validation losses for the YOLOv8 algorithm object detector and classification is observed at 100 epochs, as demonstrated on the Open Image dataset



Passageway Sequence – Camera 1

Passageway Sequence – Camera 2

Fig. 6 Sample output video frames for the passageway sequence



Camera 1 Laboratory Sequence

Camera 2 Laboratory Sequence

Fig. 7 Sample output video frames for the laboratory sequence

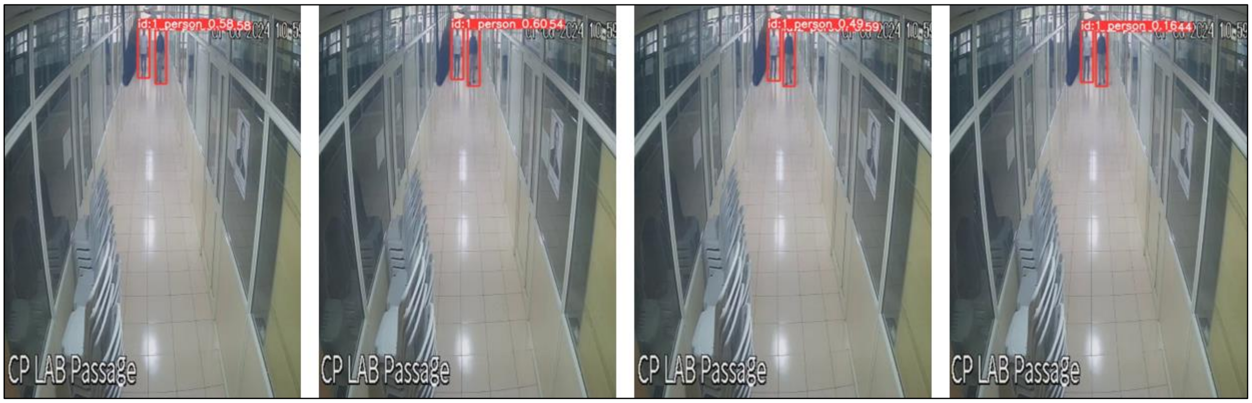


Fig. 8 Sample video of multi-camera person dataset with non-overlapping camera view - laboratory sequence

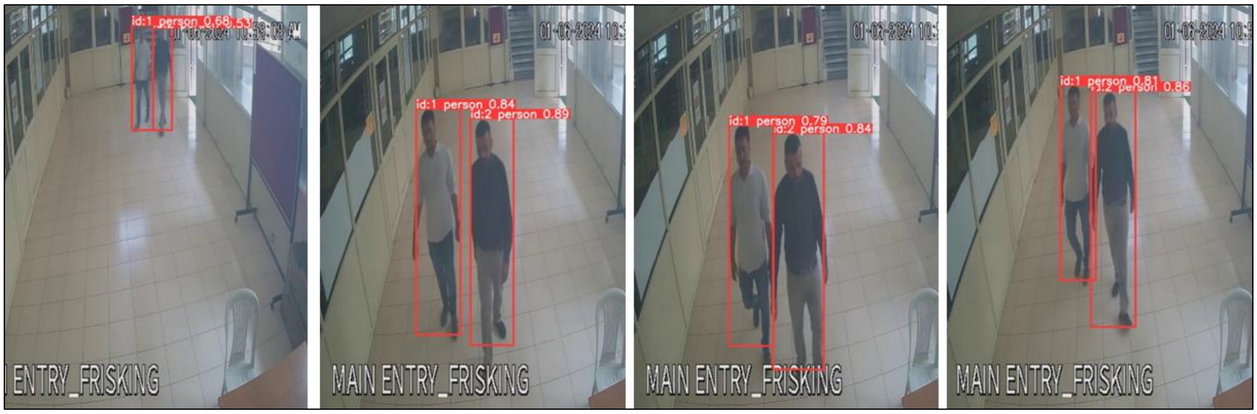


Fig. 9 Sample output of multi-camera person dataset with non-overlapping camera view - main entry

The performance of the proposed YOLOv8 and ByteTrack integration was evaluated using two challenging datasets: the EPFL Multi-camera Pedestrian Dataset (Passageway and Laboratory Sequences) and a Real-time Multi-camera Person Dataset with non-overlapping camera views. The results are presented in terms of Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP), standard metrics for tracking systems.

The MOTA for the proposed method was 70.25%, showing a slight improvement over both the K-Shortest Paths method (68%) and the Probabilistic Occupancy Map (POM) method (68%). This indicates that our approach offers more accurate tracking by correctly associating detected objects with tracks even under challenging conditions, such as poor lighting and occlusions. The MOTP of 83.3% surpasses both the K-Shortest Paths and POM methods, which achieved 82% and 70%, respectively. This demonstrates that integrating YOLOv8 and ByteTrack leads to more precise tracking object localization, even in difficult scenarios. The proposed method achieved a MOTA of 72.14% for the Laboratory Sequence, outperforming both K-Shortest Paths (70%) and POM (72%). This shows the system maintains strong tracking accuracy, even when individuals move in a controlled environment with overlapping cameras. The MOTP score for the proposed approach was 84.6%, which again improves upon the other methods (83% for K-Shortest Paths and 78% for POM). This further validates the precision of object tracking across multiple camera views in scenarios where occlusions and re-identification challenges occur frequently.

The MOTA for this dataset reached 75.60%, the highest among all evaluated scenarios. This indicates that the system excels in tracking individuals across non-overlapping camera views where maintaining identity consistency is particularly difficult. The MOTP was also impressive at 86.8%, showcasing the framework's ability to maintain high localization precision in non-overlapping camera setups, which are often more challenging due to a lack of continuous spatial information between cameras. The improvement in MOTA and MOTP metrics across all datasets demonstrates the effectiveness of integrating YOLOv8 and ByteTrack, especially in addressing occlusion, varying lighting conditions, and maintaining identity across multiple cameras.

The proposed method consistently outperforms traditional methods such as K-Shortest Paths and POM in terms of both tracking accuracy and precision, particularly in more complex environments like the EPFL Passageway and non-overlapping camera views. These results indicate that the proposed framework is suitable for tracking people in real time. It offers robust performance in challenging scenarios, making it highly applicable for intelligent surveillance systems, crowd analysis, and autonomous vehicle applications.

5. Conclusion

This paper introduces a new approach for multi-camera person tracking that leverages the strengths of YOLOv8 and ByteTrack. Through extensive experiments on challenging datasets, including the Multi-camera pedestrian video Dataset and a real-time multi-camera person dataset with non-overlapping camera views, our method has demonstrated superior performance in terms of tracking accuracy and robustness. Specifically, our approach outperformed existing state-of-the-art techniques in handling occlusions, varying lighting conditions, and complex camera perspectives.

Integrating YOLOv8 for object detection with ByteTrack's advanced tracking capabilities has proven effective in maintaining tracked individuals' identities across multiple camera views, even under challenging conditions. These results suggest that our proposed framework has significant potential for intelligent surveillance systems, autonomous vehicles, and crowd-analysis applications. Future work will explore further improvements to our method by incorporating additional data sources and enhancing the adaptability of the framework to more diverse environments.

References

[1] Saad M. Khan, and Mubarak Shah, “Tracking Multiple Occluding People by Localizing on Multiple Scene Planes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 505-519, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] O. Javed, K. Shafique, and M. Shah, “Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, vol. 2, pp. 26-33, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[3] Liliang Zhang et al., “Is Faster R-CNN Doing Well for Pedestrian Detection?,” *Computer Vision – ECCV 2016*, pp. 443-457, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[4] Joseph Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, pp. 779-788, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” *arXiv Preprint*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Yifu Zhang et al., “ByteTrack: Multi-Object Tracking by Associating Every Detection Box,” *Computer Vision - ECCV 2022*, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Nicolai Wojke, Alex Bewley, and Dietrich Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric,” *2017 IEEE International Conference on Image Processing (ICIP),* Beijing, China, pp. 3645-3649, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Alex Bewley et al., “Simple Online and Realtime Tracking,” *2016 IEEE International Conference on Image Processing (ICIP),* Phoenix, AZ, USA, pp. 3464-3468, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ivan Krasin et al., Openimages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification, 2020. [Online]. Available: https://github.com/openimages/dataset

[10] Francois Fleuret et al., “Multicamera People Tracking with a Probabilistic Occupancy Map,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267-282, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[11] Temitope Ibrahim Amosa et al., “Multi-Camera Multi-Object Tracking: A Review of Current Trends and Future Advances,” *Neurocomputing*, vol. 552, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] B. Dwyer, J. Nelson, and T. Hansen, Roboflow (Version 1.0), 2024. [Online]. Available: https://roboflow.com/research#cite

[13] Abhishek Balasubramaniam, and Sudeep Pasricha, “Object Detection in Autonomous Vehicles: Status and Open Challenges,” *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] M. Sami Zitouni et al., “Advances and Trends in Visual Crowd Analysis: A Systematic Survey and Evaluation of Crowd Modelling Techniques,” *Neurocomputing*, vol. 186, pp. 139-159, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Joseph Redmon, Ali Farhadi, “Yolov3: An Incremental Improvement,” *arXiv Preprint*, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Xingkui Zhu et al., “TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios” *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, pp. 2778-2788, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Chuyi Li et al., “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications,” *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 7464-7475, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Anton Milan et al., “MOT16: A Benchmark for Multi-Object Tracking,” *arXiv Preprint*, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[20] Laura Leal-Taixé et al., “Motchallenge 2015: Towards a Benchmark for Multi-Target Tracking,” *arXiv Preprint*, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[21] Zhong-Qiu Zhao et al., “Object Detection with Deep Learning: A Review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212-3232, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Dillon Reis et al., “Real-Time Flying Object Detection with YOLOv8,” *arXiv Preprint*, 2023. [CrossRef] [Google Scholar] [Publisher Link]