Online Multi-object Tracking with YOLOv9 and DeepSORT Optimized by Optical Flow

Djalal Djarah
Department of Electrical Engineering, University of Kasdi Merbah, Ouargla, Algeria
djarah.djalal@univ-ouargla.dz (corresponding author)

Abdeslam Benmakhlouf
Department of Electrical Engineering, University of Kasdi Merbah, Ouargla, Algeria
benmakhlouf.abdeslam@univ-ouargla.dz

Ghania Zidani
Department of Pharmacy, Mostefa Ben Boulaid University, Batna, Algeria
g.zidani@univ-batna2.dz

Laid Khettache
Department of Electrical Engineering, University of Kasdi Merbah, Ouargla, Algeria
khettache.laid@univ-ouargla.dz

Received: 19 August 2024 | Revised: 20 September 2024 | Accepted: 27 September 2024

 DOI: https://doi.org/10.48084/etasr.8770

ABSTRACT

To ensure reliable environmental perception in the realm of autonomous driving, precise and robust multi-object tracking proves imperative. This study proposes an innovative approach to multi-object tracking by combining YOLOv9's sophisticated detection capabilities with an enhanced DeepSORT tracking algorithm, enriched through the integration of optical flow. In the proposed method, the YOLOv9 detector acutely identifies objects in input images, and these detected entities are subsequently transmitted to the optimized DeepSORT tracking algorithm. The principal contribution of this study lies in improving the Kalman filter measurement model within DeepSORT by incorporating robust local optical flow, thus adding a velocity dimension to the filter's update vector. This novel approach significantly improves tracking resilience in the face of occlusions, rapid movements, and appearance changes. Evaluations on MOT17 and KITTI show substantial improvement gains of 2.42%, 2.85%, and 1.84% for HOTA, MOTA, and IDF1, respectively, on MOT17, and 1.94% in MOTA and 2.09% in HOTA on KITTI. The proposed method particularly excels in managing scenarios involving dense traffic and light variations, which are recurrent problems in dynamic urban environments. This enhanced performance positions the proposed solution as an essential component of future perception architectures for autonomous vehicles, promising safer and more efficient navigation in the complex real world.

Keywords-object detection; multi-object tracking; autonomous perception; YOLOv9; DeepSORT; optical flow

I. INTRODUCTION

Multi-Object Tracking (MOT) is a crucial issue in many modern applications, such as video surveillance [1], autonomous driving [2], and assistive robotics [3]. Recent advances in object detection have greatly stimulated the development of even more powerful MOT systems, enabling smoother interaction and a deeper understanding of dynamic environments [4]. In this context, this study proposes a new MOT approach that combines the strengths of YOLOv9 [5], a state-of-the-art object detector that stands out for its exceptional accuracy and efficiency in real-time object detection, with the DeepSORT (Deep Simple Online and Realtime Tracking) algorithm [6], renowned for its cutting-edge approach, exploiting a synergetic relationship between the Kalman filter [7] and the Hungarian algorithm [8]. This algorithm orchestrates the motion model through the Kalman filter and achieves an optimal association between observations and predicted trajectories by minimizing the overall cost of the assignment, based on a distance metric between appearance features and state predictions. Despite DeepSORT's significant performance in real-time, it encounters difficulties when object movements become more complex or when detections become intermittent. The predictions of the Kalman filter can then diverge from the real observations, generating discontinuities in the trajectories. To overcome these difficulties, a robust optical

flow [9] was integrated to provide a detailed estimate of the motion between successive images. This innovative fusion significantly improves tracking accuracy and robustness, particularly in complex scenarios characterized by occlusions, changes in appearance, and rapid movements.

Many studies have made considerable progress in the field of MOT since the integration of deep learning. Object detection, an essential preliminary stage, has benefited from the emergence of high-performance Convolutional Neural Network (CNN) architectures [10]. Single-phase detectors, such as RetinaNet [11] and CenterNet [12], have established themselves due to their efficiency and simplicity, and are widely adopted in tracking methods [13]. The YOLO series [14, 15] offers an attractive compromise between accuracy and processing speed. More recently, transformer-based architectures [16, 17] have revolutionized the field of computer vision, demonstrating a remarkable ability to model long-range relationships between different elements in a scene. Their application to tracking multiple objects [18] has opened new perspectives in managing complex and dynamic scenes. Data association in MOT is a fundamental process that begins by calculating the similarity between tracklets and detection boxes. This process then relies on various strategies to match these elements based on their similarity. Visual cues, such as position, motion, and appearance, are crucial to data matching in the MOT domain. The SORT algorithm [19] uses a Kalman filter to predict the position of tracklets in the next image and measures the Intersection over Union (IoU) between the predictions and the detection boxes as an indicator of similarity. Other more advanced approaches use networks dedicated to learning object trajectories [20], thus increasing robustness to significant camera movements or low frame rates. Position and motion similarities are particularly reliable for short-term association, while appearance similarity, quantified by the cosine similarity of the re-identification features, is essential for long-term association. An object can be re-identified thanks to appearance similarity even after prolonged occultation.

DeepSORT incorporates a separate re-identification model (Re-ID) to extract visual attributes from detection boxes. Recent models that combine object detection and re-identification [21, 22] have gained popularity due to this integrated and powerful approach. Once similarities are calculated, various matching strategies are deployed to assign identities to detected objects. These strategies can include the Hungarian algorithm or gluttonous assignment methods. The SORT system performs a single pairing between detection boxes and tracklets, while DeepSORT uses a cascading matching strategy, prioritizing recent tracklets before tackling lost ones. MOTDT [23] uses appearance similarity for the initial matching of the tracklets, followed by the use of IoU for the remaining ones. QDTrack [24] transforms appearance similarity into probabilities via a bidirectional SoftMax function and uses nearest-neighbor matching. The attention mechanism allows boxes to propagate directly in images, establishing an implicit association [25]. Innovative techniques introduce tracking queries that anticipate future positions of tracked objects without using the Hungarian algorithm, but relying on the dynamic interaction of attention mechanisms [26]. These theoretical and methodological advances in MOT improve the accuracy and robustness of object-tracking systems in complex environments.

The main contributions of this study are as follows:

- Innovative hybrid model: Presents a hybrid tracking model that takes advantage of the strengths of YOLOv9, DeepSORT, and optical flow to achieve superior tracking results.
- Improved robustness: The proposed approach is particularly effective at handling occlusions, appearance changes, and fast movements, thanks to the integration of optical flow.
- Experimental validation: The proposed method was evaluated on recognized benchmarks (MOT17 and KITTI), demonstrating its robustness against occlusions and its ability to accurately track objects in dynamic environments.

II. METHOD

The proposed MOT pipeline, shown in Figure 1, is based on a sophisticated bipartite architecture. Initially, a YOLOv9 CNN is deployed for object detection. This model, characterized by its single-pass architecture and exceptional inference speed, uses implicit convolutions and dense residual blocks to extract robust hierarchical features. YOLOv9 also incorporates a spatial attention mechanism and a multiscale prediction system, allowing the accurate detection of objects of various sizes. The network provides not only bounding boxes but also confidence scores and classifications for each detected object, with remarkable accuracy even in complex scenarios. In parallel, an optical flow estimator based on the Kanade-Lucas-Tomasi (KLT) algorithm is used to calculate the apparent displacement vectors between two consecutive images, thus providing robust kinematic information.

These two sources of information are then synergistically merged in a Kalman filter, which recursively estimates the position and velocity of each object, thus integrating the complementary advantages of the detection and tracking approaches. An optimized Hungarian assignment algorithm is then implemented to establish correspondences between current detections and Kalman filter estimates, taking into account both kinematic information (provided by the optical flow) and appearance information (extracted from visual descriptors). This innovative hybrid approach, combining the strengths of deep detection and optical flow tracking, ensures robust and accurate object tracking in complex video sequences, while mitigating the limitations inherent in each method individually. Thus, the created synergy enables effective management of partial occlusions, rapid changes in appearance, and nonlinear movements, which are recurrent problems in autonomous driving scenarios in dynamic urban environments.

A. The YOLOv9 Detector

The YOLOv9 detector, shown in Figure 2, represents a major advance in real-time object detection, challenging the current limits of computer vision.

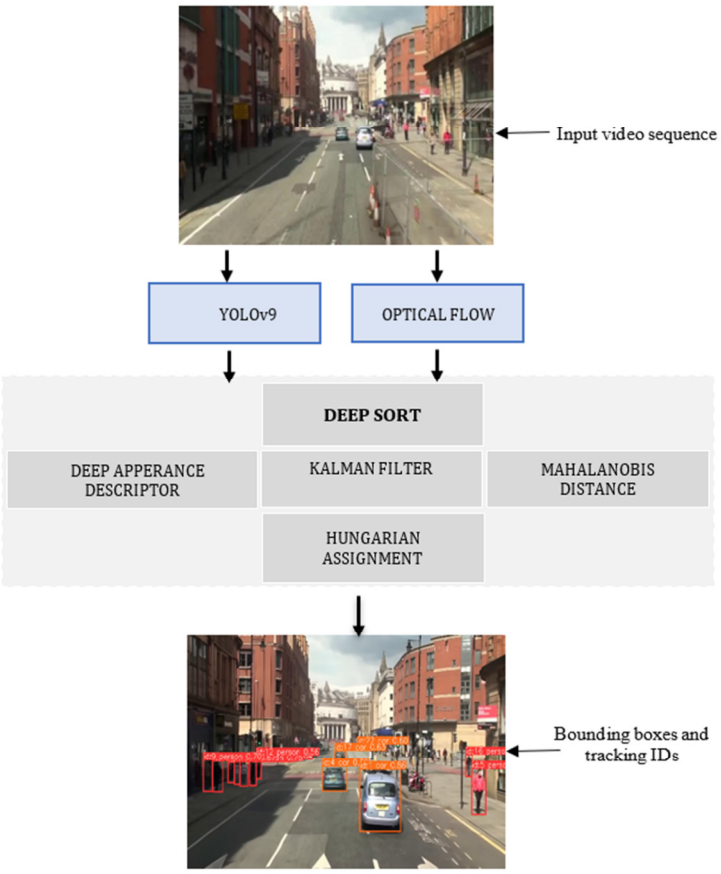

Fig. 1. The flowchart of the proposed object tracking algorithm with YOLOv9 and DeepSORT optimized with optical flow.

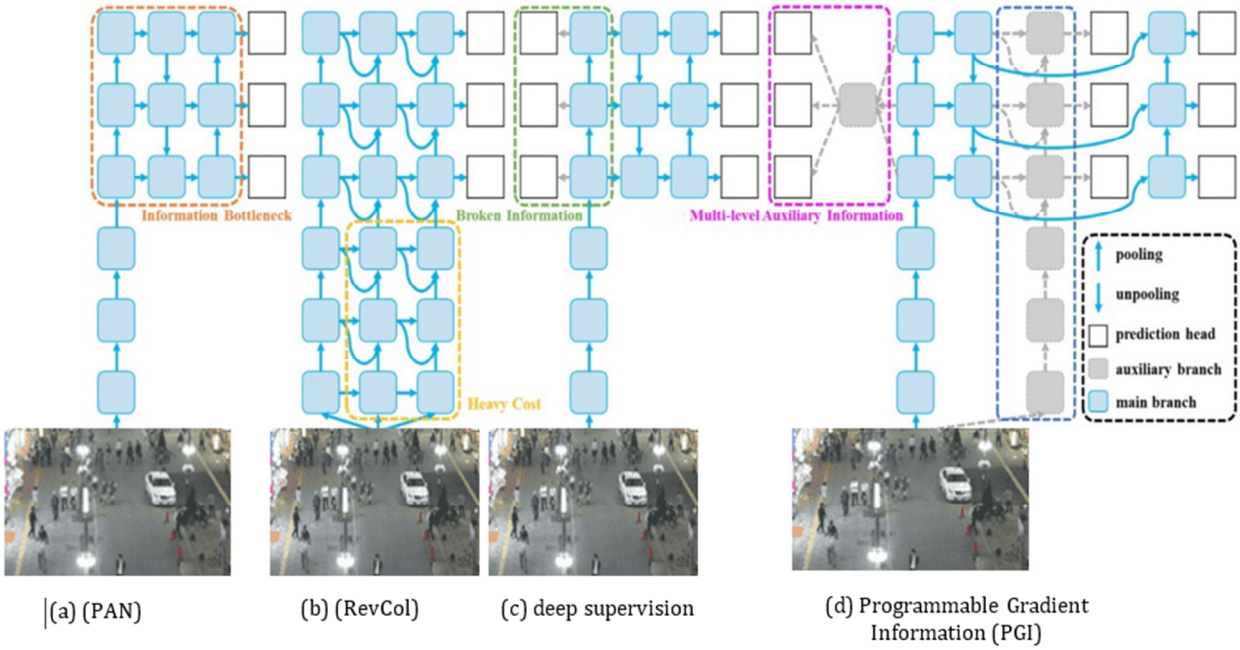

Fig. 2. PGI and related network architectures and methods [5]: (a) Path Aggregation Network (PAN), (b) Reversible Columns (RevCol) [3], (c) conventional deep supervision, and (d) the proposed Programmable Gradient Information (PGI). PGI is mainly composed of three components: (i) main branch: architecture used for inference, (ii) reversible auxiliary branch that generates reliable gradients to feed the main branch for backward transmission, and (iii) multi-level auxiliary information that controls learning of the plannable main branch at several levels of semantic information.

Its innovative architecture minimizes information loss in deep neural networks through dense connections and implicit knowledge, maintaining rich multiscale feature maps. An advanced attention mechanism and anchorless sensing head optimize accuracy and efficiency, particularly in complex scenarios with occlusions and dense objects. YOLOv9's algorithmic advances, building on the foundations laid by its predecessors and exploring new frontiers, enable it to achieve an optimal balance between average accuracy (mAP) and computational efficiency. While retaining exceptional inference speed, the results obtained on the COCO dataset [5] reveal a significant improvement in object detection performance with YOLOv9. By achieving an optimal balance between accuracy and efficiency, the YOLOv9 models outperform their predecessors, notably YOLOv8 [27]. These advances are attributable to improvements in model architecture, such as the introduction of flexible aggregation blocks, as well as innovative training strategies.

B. DeepSORT Algorithm

DeepSORT represents a significant advance in the field of MOT, offering a robust and accurate solution for the continuous localization and identification of objects in complex video sequences. This approach combines two complementary sources of information, which are then merged within a Kalman filter, providing a more accurate estimate of the dynamic state of the objects being tracked. Building on the SORT framework, DeepSORT introduces sophisticated data association mechanisms that exploit both kinematic information and object appearance. At the heart of DeepSORT is a deep feature extraction network, designed to capture discriminative visual representations that are invariant to geometric and photometric transformations. These visual descriptors, combined with a Kalman filter-based motion model, enable matches to be established between successive detections of the same object, even in the presence of temporary occlusions or variations in appearance. DeepSORT's cascade association phase is based on a multimodal similarity metric that incorporates both a Mahalanobis distance to measure the kinematic compatibility of trajectories and a cosine distance to assess the similarity of visual descriptors. This approach effectively manages ambiguities linked to similar appearances or rapid movements. Finally, DeepSORT integrates trajectory life management mechanisms, allowing traces to be created, updated, and deleted according to their consistency and reliability. These mechanisms contribute to the tracking robustness and the quality of the results obtained.

C. Kalman Filter

The use of the Kalman filter for kinematic modeling of objects in a visual tracking context can be formalized by considering an octo-dimensional state vector defined as follows

$$S = (x, y, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})^T \quad (1)$$

where (x, y) represents the coordinates of the bounding box centroid, γ denotes the aspect ratio, and h is the height. The components $(\dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$ correspond to the respective time derivatives, characterizing the kinematics of the object in the image plane.

The Kalman filter, based on the assumption of linearity and Gaussianity, operates in two distinct phases: prediction and updating. In the prediction phase, the a priori state is estimated according to

$$\hat{S}_k^- = FS_{k-1} + w_{k-1} \quad (2)$$

where F is the state transition matrix and w_{k-1} represents the process noise. The update phase incorporates z_k observations to refine the estimate

$$\hat{S}_k = \hat{S}_k^- + K_k(z_k - H\hat{S}_k^-) \quad (3)$$

where H is the observation matrix and K_k is the optimal Kalman gain. This rigorous mathematical formulation of the Kalman filter, coupled with an effective association strategy, enables accurate and computationally efficient modeling of object kinematics in a multi-object tracking framework.

D. Improved DeepSORT network

This study proposes an extension of the DeepSORT framework to improve the robustness and accuracy of MOT by integrating optical flow estimation. This hybrid approach combines the advantages of CNNs for discriminative visual feature extraction and classical optical flow tracking methods for accurate motion estimation. Initially, a YOLOv9 network is used to generate object detections and associated bounding boxes. In parallel, a KLT optical flow estimator is used to calculate apparent displacement vectors between two consecutive images. These two complementary sources of information are then merged within an extended Kalman filter, providing a more accurate estimate of the dynamic state of the objects being tracked.

E. Datasets and Metrics

1) Datasets

The MOT17 package [28] has substantially enriched the field of investigation of real-time MOT by introducing a corpus of stabilized videos acquired using various perceptual modalities. Comprising seven sequences dedicated to the model learning and validation phases, this benchmark has also consolidated its normative character by integrating the predictions of standard detectors such as DPM, Faster R-CNN [29], and SDP [30]. The KITTI dataset [31] has been designed to empower automotive platforms by focusing on detecting and tracking entities in road environments from onboard streams. To this end, it offers annotations in 2D and 3D modalities with an unprecedented level of precision. The relevant hybridization of KITTI's specificities with MOT17's metrological standards has generated reference frameworks that are particularly well suited to the challenges posed by multi-object localization in the singular context of intelligent driving. The KITTI tracking benchmark, based on 21 training sequences and 29 validation sequences, has become an essential tool for evaluating algorithmic paradigms developed in the related field of embedded computer vision. These databases provide a comprehensive scientific basis for testing and perfecting practical vision systems, particularly in the critical fields of driver automation and travel safety.

2) Metric Evaluation

Rigorous performance evaluation of detection and tracking systems is essential in the field of computer vision. To quantify the effectiveness of the proposed tracking algorithm, specific metrics were used, including:

- IDF1 criterion [32]: The F1 identification index (*IDF1*) is a composite measure that reflects the overall quality of the association between detections and pedestrian identities over time. It corresponds to the harmonic mean of identification precision (*IDP*) and identification recall (*IDR*), providing a balanced assessment of the system's ability to correctly assign a unique identity to each tracked object.

$$IDF1 = \frac{2|IDTP|}{2|IDTP|+|IDFP|+|IDFN|} \quad (4)$$

- MOTA (Multiple Object Tracking Accuracy): This fundamental metric quantifies the overall accuracy of MOT by integrating false positives, false negatives, and identity change errors, as

$$MOTA = 1 - \frac{\sum_{t=1}^{/t/}(FN_t+FP_t+IDs_t)}{\sum_{t=1}^{/t/} GT_t} \quad (5)$$

  where Ground Truth (*GT*), False Positives (*FP*), Identity Changes (*IDs*), and False Negatives (*FN*) represent real objects, false detections, identification errors, and tracker omissions, respectively.

- HOTA (Higher Order Tracking Accuracy) [33] is an advanced metric for evaluating tracking accuracy in computer vision. It excels at distinguishing superimposed objects and detecting incoherent movements, incorporating the probability of transition errors. It is given by

$$HOTA = \frac{TP-Mismatches-FT}{TP+FP+FN-Mismatches-FT} \quad (6)$$

  where *TP*, FP, FN, *Mismatches*, and *FT* represent true positives, false positives, false negatives, tracking errors, and false transitions between objects, respectively. HOTA simultaneously evaluates object detection, association, and localization, offering a more comprehensive analysis than IDF1 or MOTA. This holistic metric takes into account localization, association, and fragmentation errors for an in-depth evaluation of tracking performance.

III. EXPERIMENTAL RESULTS AND ANALYSIS

This study proposes a unified experimental framework for the comparative evaluation of object detection and tracking methods. By implementing an optimized implementation of DeepSORT that exploits optical flow and comparing it with the conventional DeepSORT implementation, an in-depth performance analysis was performed on MOT17 and KITTI benchmarks. Combining these approaches with the YOLOv9 and YOLOv8 detectors aims to provide a comprehensive and rigorous assessment of MOT capabilities under a variety of conditions.

A. Quantitative Analysis

The results presented in Table 1 demonstrate the systematic superiority of the proposed tracking algorithm over DeepSORT for all the sequences evaluated. In particular, the HOTA metric, which simultaneously quantifies localization accuracy and identity association, reveals an average improvement of 2.42% for the proposed approach, with scores ranging from 22.55% to 40.18%, compared to the range of 19.75% to 37.98% observed using DeepSORT. These results attest to a substantial improvement in the proposed algorithm's ability to accurately track objects and establish consistent identity associations. Furthermore, MOTA scores, which measure the prevalence of false positives and false negatives, corroborate this trend, with an average improvement of 2.85%. This metric, crucial for assessing the overall robustness of the tracking system, underscores the increased effectiveness of the proposed approach in minimizing detection and tracking errors. Finally, the IDF1 metric, which specifically assesses the quality of identity association across the entire sequence, shows an average increase of 1.84%. This significant improvement testifies to the increased robustness of the proposed method, particularly regarding long-term follow-up and the management of temporary occlusions.

To assess the impact of detector choice on the performance of the proposed tracking algorithm, experiments were carried out by substituting YOLOv9 for YOLOv8. The results shown in Table II corroborate the superiority of the proposed approach over DeepSORT, irrespective of the experimental configuration adopted. The HOTA, MOTA, and IDF1 scores show that the proposed MOT algorithm is less sensitive to detector changes than the conventional DeepSORT algorithm. In particular, when switching from YOLOv9 to YOLOv8, performance degradation is much greater for DeepSORT than for the proposed approach. For the HOTA score, the drop is 1.33% for DeepSORT versus only 0.43% for the proposed algorithm. Therefore, the proposed algorithm is better able to adapt to slight variations in detection quality. MOTA decreases by 1.05% for DeepSORT, but only by 0.2% for the proposed method. As MOTA directly evaluates tracking errors, the proposed approach is more robust to detector inaccuracies. The difference is similar for IDF1, which evaluates the quality of target identification over time (a drop of 1.04% for DeepSORT versus only 0.43% for the proposed algorithm). These results significantly demonstrate the greater resilience of the proposed framework to the qualitative hazards inherent in real detections. DeepSORT appears to be more coupled to the underlying detector, whereas the proposed solution seems to take advantage of the detections more efficiently.

Therefore, it can be concluded that the proposed approach, by further integrating the specificities of detections into the tracking process, demonstrates increased robustness, which is essential in complex real-world application contexts. The results in Table III demonstrate the systematic superiority of the proposed algorithm over DeepSORT on all the sequences in the KITTI benchmark. On average, the proposed algorithm achieved a 2.09% improvement in HOTA over DeepSORT. More specifically, on pedestrian sequences, the proposed algorithm achieved a HOTA score of 40.39%, an improvement

of 2.2% over DeepSORT. Similarly, for vehicle sequences, an average HOTA score improvement of 1.97% was observed. These performance gains are also notable for the MOTA and IDF1 metrics, with average improvements of 1.94% and 3.23%, respectively. These quantitative results underline the robustness and superior efficiency of the proposed approach, particularly in the complex scenarios of the KITTI benchmark.

TABLE I. TRACKING PERFORMANCE ON THE KITTI BENCHMARK WITH YOLOV9

Sequence	Trackers	Sequences of MOT17 Dataset							
		HOTA↑	MOTA↑	IDF1↑	MT↑%	ML↓%	FP↓	FN↓	IDs↓
MOT17 -01	DeepSORT	21.30	31.60	36.01	19.57	39.20	63	6125	18
	Proposed	22.55	33.44	36.98	27.33	36.13	33	5012	10
MOT17 -04	DeepSORT	37.83	49.60	55.41	21.30	34.38	1730	53911	207
	Proposed	38.97	51.09	55.11	26.20	30.17	1529	49027	198
MOT17 -06	DeepSORT	37.98	50.30	57.12	22.30	31.30	449	5992	79
	Proposed	39.01	52.14	61.33	30.47	28.11	326	5211	109
MOT17 -09	DeepSORT	26.33	37.19	43.10	19.50	32.09	329	10779	101
	Proposed	30.98	42.21	46.03	28.71	30.51	401	10121	92
MOT17 -11	DeepSORT	37.01	48.65	59.21	23.75	30.75	211	3729	33
	Proposed	40.18	53.02	59.87	41.03	26.33	197	3631	28
MOT17 -13	DeepSORT	19.75	30.65	38.55	18.41	27.09	115	13012	100
	Proposed	23.02	33.20	41.14	21.09	24.16	95	12883	76
MOT17-Mean	DeepSORT	30.03	41.33	48.23	20.80	32.46	2897	93548	558
	Proposed	**32.45**	**44.18**	**50.07**	**29.13**	**29.22**	**2581**	**85885**	**513**

TABLE II. TRACKING PERFORMANCE ON THE MOT17 BENCHMARK WITH YOLOV8

Sequence	Trackers	Sequences of MOT17 Dataset							
		HOTA↑	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
MOT17 -01	DeepSORT	19.348	29.98	34.65	18.02	41.30	69	6303	23
	Proposed	21.89	32.96	35.86	26.87	36.96	34	5015	10
MOT17 -04	DeepSORT	36.03	47.95	54.31	20.55	36.88	1789	53998	225
	Proposed	38.13	50.99	54.89	25.59	30.98	1538	49089	203
MOT17 -06	DeepSORT	36.85	49.56	55.98	21.60	32.48	462	6002	82
	Proposed	38.94	51.89	61.04	30.07	29.06	334	5225	113
MOT17 -09	DeepSORT	25.12	36.22	41.96	19.05	33.17	335	10788	109
	Proposed	30.36	41.95	45.89	27.89	31.21	417	10139	96
MOT17 -11	DeepSORT	36.12	48.45	58.29	22.82	31.84	221	3745	39
	Proposed	40.02	52.91	59.23	40.79	26.89	199	3632	28
MOT17 -13	DeepSORT	18.79	29.56	37.99	17.10	29.03	126	13089	107
	Proposed	22.83	32.89	40.96	20.93	23.67	97	12886	76
MOT17-Mean	DeepSORT	28.70	40.28	47.19	19.85	34.11	3002	93925	585
	Proposed	**32.02**	**43.93**	**49.64**	**28.69**	**29.75**	**2619**	**85986**	**526**

TABLE III. TRACKING PERFORMANCE ON THE KITTI BENCHMARK WITH YOLOV9

Metrics	Trackers	Kitti Sequence-16 Person	Kitti Sequence-18 Person	Kitti Sequence-10 Vehicle	Kitti Sequence-17 Vehicle	Kitti Mean
HOTA↑	DeepSORT	40.17	36.21	51.31	54.30	45.49
	Proposed	41.65	39.13	52.82	56.73	**47.58**
MOTA↑	DeepSORT	56.20	48.07	55.11	58.21	54.39
	Proposed	58.31	51.77	56.30	58.95	**56.33**
IDF1↑	DeepSORT	65.30	69.95	70.02	73.11	69.09
	Proposed	69.43	73.03	72.54	74.31	**72.32**

B. Qualitative Analysis

Figure 3 presents visualization results of complex cases that the proposed tracker handled effectively. Three MOT17 sequences and one KITTI sequence were selected to illustrate complex cases. The complex situations include occlusion (MOT17-04, MOT17-06, MOT17-09), motion blur (MOT17-06, KITTI-04) and the presence of small objects (MOT17-09). In the MOT17-04 video sequence, two people stand out for their behavior when it comes to keeping their ID. Person ID09 showed remarkable stability by keeping his ID even when moving back and forth. This consistency is essential for accurate tracking of people in the video sequence.

ID02 also retained its identity after an occlusion, demonstrating the robustness of the algorithms in the face of visual disturbances. After 428 images, its identifier was still valid. In the MOT17-06 video sequence, the object bearing ID85 (image #0376) completely exits the camera field in frame #0397. It reappears in frame #0415 with a change of direction while retaining the same identity. This continuity of identification, despite leaving the field of view and changing direction, illustrates the ability of the proposed algorithm to maintain a reliable and consistent association of identifiers, even under complex tracking conditions. In the MOT17-09 video sequence, the objects identified by ID020 and ID021 change direction and adopt various positions, including

variations in size. Despite these challenges, including occlusions, these objects retain their identifiers from frame #0177 to frame #0371. This algorithmic robustness demonstrates the effectiveness of the proposed method in maintaining persistent identification consistency in complex MOT scenarios, characterized by significant kinematic and morphological variations in the targets. The KITTI-04 sequence in particular illustrates this performance: in image #0003, the vehicular entity identified by ID05 retains its unique identity marker despite being completely occluded by object ID01 in image #0006.

Fig. 3. Quantitative evaluation of MOT on the KITTI and MOT17 benchmarks.

Figure 4 presents example tracking sequences to visually evaluate the performance of the proposed approach, showing a direct comparison between the results obtained using the proposed DeepSORT method enriched with optical flow and those obtained using the conventional DeepSORT method. The comparative analysis highlights the superiority of the proposed method combining DeepSORT and optical flow over the standard DeepSORT pipeline, regarding several key evaluation criteria. The red bounding boxes correspond to the proposed DeepSORT method coupled with optical flow, and the yellow bounding boxes correspond to the standard DeepSORT. The proposed method provides bounding boxes that better match the morphology of the target, even in conditions of high visual clutter (image #0177 MOT17-11-DPM).

Fig. 4. Visual comparison of MOT performance between standard DeepSORT and the proposed DeepSORT combined with optical flow.

Partial occultations are managed due to constant delimitation (image #0177, MOT17-11-DPM). The ability to adapt to variations in scale is due to refining the bounding boxes as a function of distance from the object (image #0406, MOT17-13-SDP). The robustness in the face of visual ambiguities is illustrated by constant boxes under changing lighting (image #0406 MOT17-13-SDP). This phenomenon highlights the system's ability to effectively manage temporary occlusions thanks to advanced trajectory prediction and robust re-identification mechanisms, ensuring tracking continuity even in the momentary absence of direct visual information. Finally, Figures 3 and 4 effectively illustrate the ability of the proposed algorithm to operate in various traffic scenarios, characterized by a wide range of lighting conditions, variable object sizes, and different levels of clutter and obstruction. These examples highlight the robustness and effectiveness of the proposed tracking approach in complex environments, consolidating its potential for critical applications such as autonomous driving.

IV. CONCLUSION AND OUTLOOK

This research was motivated by the persistent challenges of MOT in dynamic urban environments, where existing methods struggle to maintain high accuracy and consistency in the face of frequent occlusions and rapid scene changes. This study identified significant gaps in the integration of state-of-the-art detectors with tracking algorithms, as well as in the effective exploitation of optical flow information to improve tracking robustness. This innovative work fills these gaps through the synergistic integration of YOLOv9, the state-of-the-art in object detection, with an optimized version of DeepSORT. The innovative incorporation of optical flow into the DeepSORT architecture significantly improves tracking consistency during rapid movements or partial occlusions. The domain-specific contributions of this study include the following:

- A quantifiable improvement in tracking performance, with gains in HOTA, MOTA, and IDF1 on the KITTI and MOT17 benchmarks, compared to conventional DeepSORT.
- A new evaluation framework for tracking robustness in complex urban scenarios, including metrics for tracking stability during prolonged occlusions.
- A modular and extensible architecture that facilitates the future integration of additional perception modalities and adaptation to various operational environments.

This study sets a new benchmark for perception systems in autonomous vehicles, paving the way for more reliable and safer applications in complex urban environments. Future work will focus on integrating advanced predictive capabilities and extending this approach to multi-sensor scenarios, promising significant advances in real-time autonomous perception.

NOMENCLATURE

Multi-Object Tracking (MOT): The core concept of tracking multiple objects in a scene.

You Only Look Once version 9 (YOLOv9): A specific object detection algorithm, particularly its v9 variant.

Deep Simple Online and Realtime Tracking (DeepSORT): A popular multi-object tracking algorithm that combines a deep appearance model with a Kalman filter.

Convolutional Neural Network (CNN): A neural network type commonly used for image and video analysis.

Intersection over Union (IoU): A metric used to measure the overlap between two bounding boxes.

Re-identification model (Re-ID): A model used to match an object in one image with the same object in another image.

Kanade-Lucas-Tomasi (KLT): A feature tracking algorithm that is often used in computer vision.

F1 identification index (IDF1): A metric used to evaluate the performance of re-identification models.

Multiple Object Tracking Accuracy (MOTA): A popular metric to evaluate MOT algorithms.

Higher Order Tracking Accuracy (HOTA): A comprehensive metric that considers tracking accuracy, identification accuracy, and false positives/negatives.

REFERENCES

[1] V. Saikrishnan and M. Karthikeyan, "Mayfly Optimization with Deep Learning-based Robust Object Detection and Classification on Surveillance Videos," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11747–11752, Oct. 2023, https://doi.org/10.48084/etasr.6231.

[2] F. Dang, D. Chen, J. Chen, and Z. Li, "Event-Triggered Model Predictive Control With Deep Reinforcement Learning for Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 459–468, Jan. 2024, https://doi.org/10.1109/TIV.2023.3329785.

[3] X. Sun, X. Weng, and K. Kitani, "When We First Met: Visual-Inertial Person Localization for Co-Robot Rendezvous," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, NV, USA, Oct. 2020, pp. 10408–10415, https://doi.org/10.1109/IROS45743.2020.9341739.

[4] 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, https://doi.org/10.1109/TPAMI.2016.2577031.

[5] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information." arXiv, Feb. 28, 2024, https://doi.org/10.48550/arXiv.2402.13616.

[6] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, Sep. 2017, pp. 3645–3649, https://doi.org/10.1109/ICIP.2017.8296962.

[7] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960, https://doi.org/10.1115/1.3662552.

[8] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955, https://doi.org/10.1002/nav.3800020109.

[9] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, Vancouver, Canada, Dec. 1981, vol. 2, Art. no. 674–679, [Online]. Available: https://hal.science/hal-03697340.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, Oct. 2020, https://doi.org/10.1109/TPAMI.2018.2844175.

[11] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, https://doi.org/10.1109/TPAMI.2018.2858826.

[12] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as Points." arXiv, Apr. 25, 2019, https://doi.org/10.48550/arXiv.1904.07850.

[13] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, and H. Lu, "Improving Multiple Object Tracking with Single Object Tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 2453–2462, https://doi.org/10.1109/CVPR46437.2021.00248.

[14] T. Saidani, "Deep Learning Approach: YOLOv5-based Custom Object Detection," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12158–12163, Dec. 2023, https://doi.org/10.48084/etasr.6397.

[15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO Series in 2021." arXiv, Aug. 05, 2021, https://doi.org/10.48550/arXiv.2107.08430.

[16] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards Real-Time Multi-Object Tracking," in *Computer Vision – ECCV 2020*, Glasgow, UK, 2020, pp. 107–122, https://doi.org/10.1007/978-3-030-58621-8_7.

[17] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu, "Rethinking the Competition Between Detection and ReID in Multiobject Tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 3182–3196, 2022, https://doi.org/10.1109/TIP.2022.3165376.

[18] D. Meng *et al.*, "Conditional DETR for Fast Training Convergence," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, Oct. 2021, pp. 3631–3640, https://doi.org/10.1109/ICCV48922.2021.00363.

[19] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 3464–3468, https://doi.org/10.1109/ICIP.2016.7533003.

[20] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe, "SiamMOT: Siamese Multi-Object Tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 12367–12377, https://doi.org/10.1109/CVPR46437.2021.01219.

[21] 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, https://doi.org/10.1007/s11263-021-01513-4.

[22] X. Zhou, T. Yin, V. Koltun, and P. Krahenbuhl, "Global Tracking Transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8761–8770, https://doi.org/10.1109/CVPR52688.2022.00857.

[23] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6, https://doi.org/10.1109/ICME.2018.8486597.

[24] J. Pang *et al.*, "Quasi-Dense Similarity Learning for Multiple Object Tracking," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 164–173, https://doi.org/10.1109/CVPR46437.2021.00023.

[25] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Aug. 01, 2023, http://doi.org/10.48550/arXiv.1706.03762.

[26] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "TrackFormer: Multi-Object Tracking with Transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8834–8844, https://doi.org/10.1109/CVPR52688.2022.00864.

[27] Y. Ma and Z. Zhou, "Adversarial Attacks on Adversarial Bandits." arXiv, Jan. 29, 2023, https://doi.org/10.48550/arXiv.2301.12595.

[28] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking." arXiv, May 03, 2016, https://doi.org/10.48550/arXiv.1603.00831.

[29] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, Jun. 2008, pp. 1–8, https://doi.org/10.1109/CVPR.2008.4587597.

[30] F. Yang, W. Choi, and Y. Lin, "Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2129–2137, https://doi.org/10.1109/CVPR.2016.234.

[31] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013, https://doi.org/10.1177/0278364913491297.

[32] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking," in *Computer Vision – ECCV 2016 Workshops*, Amsterdam, The Netherlands, 2016, pp. 17–35, https://doi.org/10.1007/978-3-319-48881-3_2.

[33] J. Luiten *et al.*, "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 548–578, Feb. 2021, https://doi.org/10.1007/s11263-020-01375-2.