Memory-Driven Person ReID for Identity Consistency in Multi-Object Tracking

Tista Pal, Trinh Quoc Nguyen and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, I*wate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
Email: s231x018@s.iwate-pu.ac.jp, g236v201@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—Many traditional Multi-Object Tracking methods primarily emphasize detection accuracy and short-term trajectory continuity, often overlooking long-term identity consistency, which is crucial for robust person Re-Identification (ReID). This paper presents a ReID focused Multi Object Tracking (MOT) framework designed to improve long-term identity preservation through memory-driven embedding refinement rather than detector-centric enhancements. The proposed framework focuses on a ReID-focused MOT framework aimed at improving identity preservation across extended temporal spans. The framework integrated a You Only Look Once (YOLO)-based detector, DeepSORT for motion-aware association, and a Global Identity Memory module that maintains and refines identity embedding over time through memory driven fusion. In addition, a Filtered IDF1 metric is proposed to evaluate identity consistency by focusing solely on detected instances, providing a fairer assessment of long-term identity retention. To investigate the impact of feature extraction quality, representative backbones, OSNet and ResNet50, are evaluated independently under identical MOT17 benchmark conditions. Experimental results demonstrate that the proposed GlobalID framework consistently improves identity retention across different feature extractors, demonstrating that segmentation based embedding refinement combined with memory driven fusion effectively enhances robust, identity-consistent tracking in surveillance and autonomous systems.**

Keywords-Person ReID; Multi-Object Tracking; Identity Consistency; GlobalID; ResNet50; OSNet; DeepSORT.

I. INTRODUCTION

Person tracking with Re-IDentification (ReID) aims to maintain consistent identity labels for individuals across video frames, and across camera views, forming a crucial component in surveillance systems, public safety systems and intelligent transportation systems [1]. In this work, we focus on single camera MOT, where the goal is to detect individuals and preserve their identities over time within a single continuous video stream.

Maintaining identity consistency remains challenging due to real-world visual variability. Changes in illumination, pose, orientation, occlusion and camera motion can break the appearance continuity of a person, leading to misidentification when they reappear after temporary disappearance [2]. Moreover, individuals who are wearing similar clothes introduce ambiguity in feature associations, often resulting in ID switches or identity merging [3]. These failure cases highlight the need for robust appearance modelling and reliable identity preservation mechanisms in single camera MOT systems.

Many traditional tracking systems, including Kalman filter-based motion models [4] and appearance-augmented approaches, such as SORT [5] and DeepSORT [6] have made progress in reducing fragmented trajectories. However, these methods typically rely on frame-wise feature matching, lacking mechanisms to maintain long-term identity memory or adaptively update identity embeddings. Consequently, they struggle to preserve stable identity assignments under dynamic conditions. Recent research has shown that deep feature embeddings play a crucial role in identity consistency [7]. Networks, such as ResNet50 [8] and Omni-Scale Network (OSNet) [9] have demonstrated strong discriminative power in static ReID benchmarks, but their performance within continuous tracking scenarios remains underexplored.

To address these limitations, this study focuses on improving long-term identity consistency in ReID-enhanced multi-object tracking. We hypothesize that the stability and discriminative capacity of feature embeddings are key factors in maintaining consistent identity assignment over time. Accordingly, our framework introduces three key innovations. First, we employ robust person detection and appearance feature extraction to obtain identity embeddings suitable for temporal refinement. Second, we propose a Global Identity Memory (GlobalID) module that persistently updates identity embeddings using Exponential Moving Averages (EMA), inspired by memory-based representation learning in unsupervised ReID [10], enabling adaptation to gradual appearance changes and reducing identity fragmentation. Third, we integrate cosine similarity, EMA-based embedding fusion, and Intersection over Union (IoU)-based motion cues into a unified association cost to improve robustness under occlusion and motion drift.

Additionally, we investigate segmentation-guided embedding refinement as an auxiliary enhancement to analyze its impact on identity preservation. Experimental results indicate that comparable identity consistency can be achieved without segmentation when memory-driven fusion is applied, highlighting the dominant role of global identity memory in long term tracking.

To more accurately evaluate identity stability in tracking, we introduce a Filtered IDF1 metric, designed to isolate the effect of identity association mechanisms from detection errors. Unlike the standard IDF1 score, which is influenced by missed detections and false positives, the proposed metric evaluates only the successfully detected instances, providing a more interpretable measure of temporal identity consistency. Comprehensive experiments on the MOT17 benchmark, we demonstrate that these enhancements effectively reduce ID switching and surpass the accuracy of conventional DeepSORT-based systems.

The remainder of this paper is organized as follows. Section II reviews related work on multi-object tracking, person re-identification, and memory-based tracking methods. Section III describes the proposed framework, including the object detection, feature extraction, and GlobalID module. Section IV presents experimental results and evaluation on the MOT17-Scale-Dependent Pooling (SDP) benchmark. Section V concludes the paper and outlines directions for future work.

II. RELATED WORK

A. Evolution of Multi-Object Tracking

Early MOT primarily relied on handcrafted features and motion-based prediction models, such as Kalman filters [4] and optical flow [11] to estimate object trajectories across frames. While computationally efficient, these approaches were highly sensitive to occlusion, camera motion, and appearance changes, often producing fragmented trajectories and frequent identity switches which reduced their reliability in crowded or dynamic scenes.

The introduction of Simple Online and Real-time Tracking (SORT) [5] and later DeepSORT [6] marked a significant milestone in the evolution of MOT. SORT employed Kalman filtering with bounding-box IoU association, achieving impressive speed but limited identity preservation. DeepSORT improved upon this by integrating a deep appearance descriptor trained for ReID, enabling the tracker to associate detections using both motion and visual similarity. This enhancement substantially reduced ID switches and improved long-term association stability.

In parallel, object detection frameworks such as YOLO based architecture [12] [13] have gained popularity for their high accuracy and real time performance. This framework achieves high precision and frame-rate efficiency through single-stage detection pipelines. Segmentation variants further improved localization by generating pixel-level masks that effectively suppress background interference and improving feature extraction for ReID integration. However, segmentation incurs additional computational cost and its effectiveness in improving long-term identity preservation within MOT pipelines remains an open research question.

B. Advances in ReID

ReID plays a central role in enhancing tracking reliability by providing appearance-based cues for identity matching. Early Convolutional Neural Network (CNN)-based ReID models, such as ResNet-50 [8], focused on learning global appearance features, offering a robust baseline for visual representation. However, such models often struggled with fine-grained local variations, such as changes in pose or partial occlusion. Recent architecture has introduced multi-scale or part-aware learning to overcome these limitations. OSNet [9] efficiently captures both local fine-grained and global structural information, achieving strong performance with minimal computational overhead and making it suitable for real-time tracking contexts. Other studies, such as PCB (Part-based Convolutional Baseline) [3] and Multi-Granularity Network (MGN) [14] emphasize structured feature decomposition to better handle pose and viewpoint variations. Additionally, Transformer-based ReID models [15] have recently shown promise in modeling long-range dependencies and improving context awareness across scenes.

Despite these advancements, most ReID models are trained and evaluated in static conditions (e.g., Market-1501, DukeMTMC, MSMT17) and are not directly optimized for temporal identity consistency within continuous video sequences. When integrated into tracking pipelines, they still face challenges with dynamic background interference, motion blur, and lighting fluctuations.

C. Embedding-Driven and Memory-Based Tracking

In recent years, research has shifted toward embedding-driven and memory-based MOT frameworks [3], [10]. These approaches accumulate temporal embeddings to maintain consistency across frames, allowing for adaptive feature matching that extends beyond immediate temporal windows. For example, Tracktor++ [16] and FairMOT [17] integrate detection and ReID into joint frameworks, improving both tracking precision and speed. Similarly, Joint Detection and Embedding (JDE) [18] introduced end-to-end training for simultaneous detection and embedding extraction, enabling efficient real-time inference.

However, even with these advances, most existing systems emphasize short-term association and lack explicit mechanisms to ensure long-term identity preservation. Few studies address global identity management, where accumulated embeddings are updated or refined dynamically over time to mitigate ID drift caused by gradual appearance changes or occlusion. Approaches such as self-adaptive galleries [19] or open-world ReID memory systems [20] have made progress toward continuous identity learning but remain limited in maintaining stable embedding representations within unified MOT pipelines.

D. Gap in Literature

While combining object detection and ReID embeddings has improved online MOT performance, existing frameworks still lack mechanisms to preserve identity coherence over longer temporal spans. Identity inconsistency typically arises when individuals undergo pose or orientation changes, partial occlusions, or lighting variations, leading to repeated ID fragmentation. Only a limited number of studies explicitly incorporate a feature embedding-based global identity memory that evolves over time and actively guides association decisions.

To address this gap, we propose an Integrated Global Identity Memory (GlobalID) that dynamically updates identity embeddings through EMA while applying similarity and IoU thresholds to ensure stable associations. Unlike conventional short-term embedding buffers, GlobalID provides a persistent, adaptive memory structure that bridges local frame-level tracking and long-term identity preservation. Implemented within a YOLOv8 + DeepSORT + OSNet/ResNet50 pipeline, the proposed framework effectively reduces ID switching and enhances overall tracking reliability in dynamic visual environments.

III. METHODOLOGY

This section describes the design of the embedding-driven, ReID-enhanced MOT framework, which emphasizes long-term identity consistency through memory-based embedding refinement and global identity management. The framework supports two operating configurations, with and without segmentation-based embedding refinement allowing systematic analysis of the impact of feature isolation versus computational efficiency.

A. Framework Overview

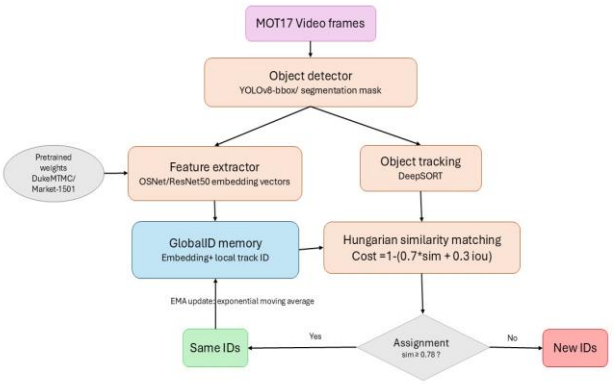

Figure 1. Proposed Framework.

The proposed framework unifies object detection, feature extraction, MOT, and ReID to study how feature quality and identity-embedding management influence long-term ID preservation. Its goal is to assess how embedding refinement and global ID strategies affect MOT performance under challenging conditions such as occlusion, illumination changes, and viewpoint variation [1].

As shown in Figure 1, the pipeline includes a YOLOv8 detector (two configurations are compared: segmentation-based setup that masks background regions before feature extraction, and a detection-based setup that uses raw bounding boxes), DeepSORT for short-term association, ReID backbones (OSNet and ResNet50) for embedding extraction, and a Global Identity Memory (GlobalID) module that maintains long-term consistency using exponential moving averages and adaptive cosine matching.

Overall, the framework provides a controlled way to analyze how embedding refinement and global identity memory contributes to stable ID assignments in realistic MOT scenarios.

B. Dataset

Experiments were conducted on the MOT17 dataset [21]-[23], which consists of multiple pedestrian video sequences captured under varying illumination, crowd density, and camera motion conditions. Each sequence provides ground-truth bounding boxes and identity annotations in the MOTChallenge format. The selected subset - MOT17-02-SDP, MOT17-04-SDP, MOT17-09-SDP, MOT17-10-SDP, and MOT17-11-SDP - these five sequences were selected to represent a diverse range of environmental and motion conditions. MOT17-02 and MOT17-04 feature static cameras with high crowd density, MOT17-09 and MOT17-11 involve low-to-medium density scenes with moderate occlusion, and MOT17-10 includes a moving camera with dynamic viewpoint shifts. Together, they provide a balanced evaluation of identity preservation under varied real-world challenges without introducing redundant or overlapping conditions.

This benchmark is widely used for MOT evaluation due to its complexity and standardization, allowing fair comparison with prior works such as DeepSORT [6] and FairMOT [17].

C. Object Detection and Segmentation

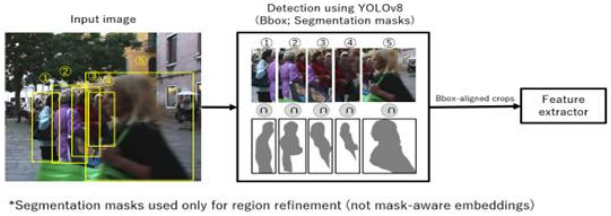

Figure 2. Person detection and segmentation using YOLOv8.

Person detection was performed using YOLOv8 architecture (Figure 2), with two configurations evaluated to study the impact of region refinement on embedding quality. For the segmentation-based configuration, YOLOv8-seg was employed to generate instance-level masks alongside bounding boxes. The segmentation masks were used to isolate person regions by suppressing background pixels before feature extraction. In contrast, the detection only configuration uses standard YOLOv8 bounding boxes directly for feature extraction, offering significantly faster inference. For both the configurations, only class 0 (person) detections were retained, using a confidence threshold of 0.3 and an IoU threshold of 0.6.

D. Feature extraction and ReID

Two representative ReID backbones, ResNet50 [8] and OSNet [9], were employed to extract feature embeddings. Each was initialized with pretrained weights from large-scale ReID datasets (Market-1501 and DukeMTMC), ensuring strong feature generalization. Person crops were resized (256×128 for OSNet, 224×224 for ResNet50) and normalized using ImageNet mean-std statistics and encoded into L2-normalized feature vectors (512-D and 2048-D, respectively)

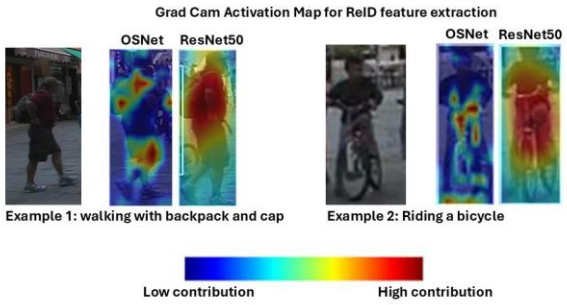

Figure 3. Activation map of features extracted by OSNet and ResNet50.

for cosine-similarity matching. These embeddings are stored in memory and on disk.

Using both backbones allows a controlled comparison between lightweight omni scale features and deeper global representations. To further examine their behavior, Grad-CAM visualization highlights the spatial regions contributing to each model's embeddings. As shown in Figure 3, OSNet attends to multiple fine-grained body regions, whereas ResNet50 focuses more on global cues such as the silhouette and torso.

E. MOT with DeepSORT

DeepSORT [6] was employed as the short-term tracking component, combining Kalman-filter-based motion prediction with appearance-based matching. Although DeepSORT effectively reduces short-term identity switches, it relies on limited temporal embedding buffers and is prone to identity drift under prolonged occlusion. In proposed framework, DeepSORT generates local track identities, which are subsequently refined and stabilized by the GlobalID module to achieve sequence-wide identity consistency. Tracking parameters used in this study are summarized in Table I.

TABLE I. DEEPSORT TRACKING PARAMETERS USED IN THIS STUDY

Parameter	Value	Description
Max_age	30	Max. No. of frames to keep a lost track alive.
nn_budget	200	Max. size of the appearance descriptor gallery.
Max_cosine_distance	0.2	Threshold for matching appearance embeddings.
Max_iou_distance	0.7	IoU threshold for matching.
n_init	3	No. of consecutive detections before confirming a new track.

F. GlobalID

To maintain identity consistency across sequences and handle occlusions or re-entries, Global Identity Memory (GlobalID) was developed. Unlike DeepSORT's limited local gallery, GlobalID functions as a persistent global memory that stores and updates identity embeddings throughout the sequence. Each new detection is compared against stored identity embeddings using cosine similarity, defined as:

$$cos(\theta) = (a \cdot b)/(\|a\|\|b\|) \quad (1)$$

Where, a and b are two feature vectors representing embeddings of the current and stored detections, respectively.

To adapt to gradual appearance changes, embeddings are updated using EMA:

$$f_t = (1-\alpha)f_{\{t-1\}} + \alpha f_{new}, \ \alpha = 0.95 \quad (2)$$

where the most recent embedding is represented by f_{new} and f_t is the updated smoothed feature vector. The chosen smoothing factor is α.

Additionally, for each identity, a maximum of 20 feature vectors were retained to prevent memory overflow and reduce noise accumulation. New identities were only confirmed after appearing consistently for 5 consecutive frames, introducing a hysteresis effect that suppresses fake ID creation.

This memory driven embedding enables the system to maintain coherent identities across long-term sequences, significantly reducing ID fragmentation and false associations compared with conventional trackers [18]-[20].

Figure 4 demonstrates the working of the GlobalID module.

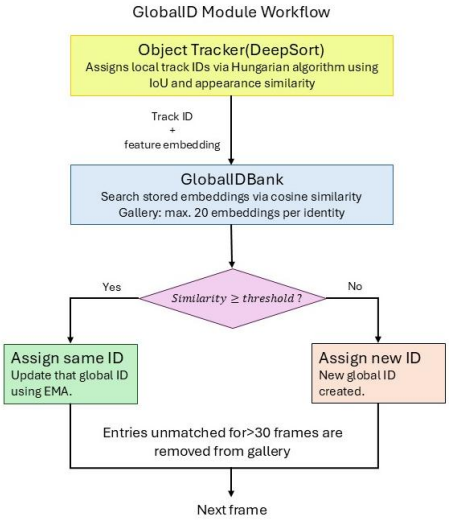

Figure 4. GlobalID Module.

IV. RESULTS

All experiments were conducted on sequences from the MOT17 dataset, which provides diverse real-world tracking scenarios involving frequent occlusions, appearance changes, and dynamic camera motion. We report both the standard MOTChallenge metric (IDF1) [22] and the proposed Filtered IDF1(F-IDF1), which evaluates identity consistency exclusively among successfully detected instances. This dual evaluation enables a clearer distinction between detection accuracy and identity-preservation capabilities.

A. Filtered IDF1 metric

Identity association performance in multi-object tracking is commonly evaluated on Association Accuracy (AssA), which measures how effectively a tracker preserves object identities across frames [22], [23]. This metric is represented by the IDF1 score, the harmonic mean of identity precision and recall, defined as:

$$F-IDF1 = (2 \times IDTP\_det) / (2 \times IDTP\_det + IDFP\_det + IDSW\_det) \quad (3)$$

where the subscript det indicates frames with valid detections, and $IDSW$ quantifies identity changes among detected objects.

This refinement provides a more focused view of ReID and memory integration effects, particularly in frameworks where detection quality is already saturated by high-performance detectors such as YOLOv8 [24].

(a) Baseline (YOLOv8 + DeepSORT) (b) The proposed method

Figure 5. The proposed method should have better consistency, not more switches.

B. *Results on MOT17-SDP*

To assess the contribution of appearance embeddings and the GlobalID module, pretrained ReID backbones (OSNet_x1_0 and ResNet50) were integrated into the YOLOv8 + DeepSORT pipeline. Each backbone was evaluated using Market-1501 [25] and DukeMTMC [22] pretrained weights, under two configurations: segmentation based embedding refinement and detection only embedding refinements. As a baseline, the standard YOLOv8 + DeepSORT configuration was evaluated using DeepSORT's built-in CNN appearance descriptor, without any external ReID backbone or GlobalID module. This represents the conventional tracking approach that our framework aims to improve upon.

Across both configurations, integrating ReID backbones with GlobalID consistently improved identity preservation on MOT17-SDP (Table II). OSNet-DukeMTMC achieved the strongest results, reaching IDF1 $\approx$ 0.39 and F-IDF1 $\approx$ 0.66 - roughly a $25-35\%$ improvement over the baseline - demonstrating that refined embeddings and global memory updates reduce ID fragmentation. F-IDF1 further shows that identity association becomes more reliable even when detection quality remains unchanged, as illustrated in Figure 5.

The segmentation assisted configurations introduced a substantial computational overhead, reducing inference speed to $1.38-2.46$ FPS. This bottleneck arises primarily from two sources, the additional forward pass required by YOLOv8-seg to generate instance masks, and the per-frame background suppression applied before feature extraction. Importantly, the ReID backbones themselves were not designed for mask-aware inputs, meaning the segmentation step adds cost without being fully exploited by the downstream feature extractors. Without segmentation, both backbones achieve significantly faster speeds ($5.73-7.13$ FPS), representing a more practical operating point for real-time applications.

TABLE II. PERFORMANCE COMPARISON ON THE MOT17-SDP BENCHMARK

Backbone	Pretrained weights	Detection Mode	Performance comparison		
			IDF1	*F-IDF1*	*FPS*
DeepSORT CNN	-	YOLOv8	0.2916	0.6412	8.51
OSNet	DukeMTMC	YOLOv8-seg	0.3918	0.6571	1.38
OSNet	Market-1501	YOLOv8-seg	0.3871	0.6583	1.44
ResNet50	DukeMTMC	YOLOv8-seg	0.3479	0.6354	2.23
ResNet50	Market-1501	YOLOv8-seg	0.3479	0.6354	2.46
OSNet	DukeMTMC	YOLOv8	0.3783	0.6514	7.13
OSNet	Market-1501	YOLOv8	0.3864	0.6531	6.55
ResNet50	DukeMTMC	YOLOv8	0.3084	0.6102	6.03
ResNet50	Market-1501	YOLOv8	0.3140	0.6203	5.73

V. CONCLUSIONS AND FUTURE WORKS

This study introduced a Re-ID-focused multi-object tracking framework that integrates YOLOv8 detection (with optional segmentation), DeepSORT association, and pretrained ReID backbones (OSNet and ResNet50) for appearance embedding. The key contribution, the GlobalID Memory module, provides a persistent, memory-driven identity refinement mechanism that maintains consistent identities across frames through EMA fusion and cosine similarity. Experiments on the MOT17-SDP benchmark demonstrate consistent improvements over the YOLOv8 + DeepSORT baseline in both IDF1 and the proposed Filtered IDF1 metric, validating the effectiveness of ReID-driven association independent of detection quality. OSNet-DukeMTMC achieved the best identity consistency (IDF1 = 0.3918, F-IDF1 = 0.6571), representing a $25-35\%$ improvement over the baseline. A few conclusions can be drawn First, segmentation-based embedding refinement improves identity consistency by suppressing background noise, but its benefit is constrained when using backbones pretrained on bounding box crops rather than mask-aware inputs, the feature extractors were not trained to exploit the cleaner segmented regions. Second, the EMA smoothing factor ($\alpha = 0.95$) proved effective for gradual appearance adaptation. Third, the gallery confirmation threshold of five consecutive frames successfully suppressed false identity

creation in crowded scenes but introduced a slight delay in registering fast moving individuals who briefly exit and re-enter the frame.

The primary limitation of the current framework is its computational cost. Segmentation assisted configurations operate at only $1.38 - 2.46$ FPS, making real-time deployment impractical. Additionally, the framework was evaluated on a subset of five MOT17-SDP sequences, and generalization to other benchmarks or camera setup remains to be verified.

Future work will focus on three directions. First, integrating mask aware ReID architectures to better exploit segmentation cues. Second, exploring lightweight architectures and embedding alternatives to reduce inference overhead to achieve real-time performance. Third, extending the GlobalID module to multi-camera settings, which represents a natural and challenging extension of the current framework.

REFERENCES

[1] W. Luo, J. Xing and X. Zhang, "Multiple object tracking: A review," arXiv preprint arXiv:1409.7618, 2014.

[2] H. Wang, S. Ullah, D. Li and Y. Liu, "Recent advances in deep learning-based person re-identification," Applied Sciences, vol. 9, no. 8, p. 1535, 2019.

[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 269-286.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, no. 1, pp. 35-45, 1960.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2016, pp. 2956-2960.

[6] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, 2017, pp.3645-3649.

[7] Z. Zhang, L. Sun, Q. Leng and S. Liao, "Towards real-time multi-object tracking with adaptive appearance models," Pattern Recognition Letters, vol. 136, pp. 213-220, 2020.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp.770-778.

[9] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-scale feature learning for person re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp.3701-3711.

[10] Y. Li, X. Zhu and S. Gong, "Unsupervised person re-identification with stochastic training strategy," IEEE Trans. Image Process., vol. 31, pp. 4240–4250, 2022.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1981, pp.674-679.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp.779-788.

[13] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[14] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2018, pp.274-282.

[15] S. He, H. Luo, P. Wang, F. Wang, H. Li and W. Jiang, "TransReID: Transformer-based object re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2021, pp.15013-15022.

[16] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking without bells and whistles," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019, pp.941-951.

[17] Y. Zhang, C. Wang, X. Wang, W. Zeng and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," Int. J. Comput. Vis., vol. 129, no. 11, p. 3069–3087, 2021.

[18] Z. Wang, L. Zheng, Y. Liu, Y. Li and S. Wang, "Towards real-time multi-object tracking," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 107–122.

[19] L. Jin, Z. Zheng and Y. Sun, "Learning a self-adaptive gallery for unsupervised person re-identification," IEEE Trans. Image Process., vol. 31, p. 5282–5294, 2022.

[20] Z. Zheng, L. Zheng and Y. Yang, "Open-world person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, p. 2630–2647, 2021.

[21] A. Milan, L. L. Taixé, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," arXiv spreprint arXiv:1603.00831, 2016.

[22] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conf. Comput. Vis. (ECCV), 2016, pp. 17–35.

[23] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. L. Taixé and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," Int. J. Comput. Vis., vol. 129, p. 548–578, 2021.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116-1124.