ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11462134

FOCALLINK: DENSELY MODULATED CONTRASTIVE LEARNING FOR TRAJECTORY ASSOCIATION IN MULTI-OBJECT TRACKING

Yanting Zhang[1,✉], *Wujian Huang*[1], *Cairong Yan*[1], *Rui Fan*[2], *Gaoang Wang*[3,✉], *and Jenq-Neng Hwang*[4]

[1]School of Computer Science and Technology, Donghua University, Shanghai, China
[2]College of Electronic and Information Engineering, Tongji University, Shanghai, China
[3] ZJU-UIUC Institute, Zhejiang University, Haining, China
[4] Department of Electrical and Computer Engineering, University of Washington, Seattle, USA

ABSTRACT

Multi-object tracking (MOT) remains challenging in complex scenes due to occlusions, detection failures, and object interactions that cause fragmented trajectories and identity ambiguity. We propose FocalLink, a trajectory association framework that enhances existing trackers, including state-of-the-art models, through three key components: (1) a temporal focal modulation mechanism that captures multi-scale temporal dependencies with linear complexity, alleviating the loss of temporal information during occlusions; (2) a densely connected architecture that preserves hierarchical features across layers, helping mitigate feature degradation under complex interactions; and (3) a contrastive learning framework that explicitly structures the embedding space for identity discrimination, thereby reducing ambiguity between similar-appearing objects. Experiments on MOT17 and MOT20 show that FocalLink consistently improves performance across HOTA, MOTA, IDF1, and ATA, while significantly reducing identity switches, demonstrating both its theoretical and practical effectiveness.

***Index Terms*—** Multi-Object Tracking, Contrastive Learning, Focal Modulation

1. INTRODUCTION

Multi-object tracking (MOT) is a fundamental computer vision task with applications in autonomous driving, intelligent surveillance, and traffic analysis. The goal of MOT is to recover complete target trajectories while maintaining identity consistency across frames. Although advances in object detection have improved per-frame accuracy, preserving object identities remains challenging in real-world scenarios involving occlusions, dense crowds, and abrupt motion changes. Such challenges often lead to identity switches, including misassigned IDs or fragmented trajectories, which significantly degrade tracking performance (Fig. 1).

Conventional tracking-by-detection frameworks split MOT into object detection and data association [1, 2], but fail to sustain identity consistency under occlusions, missed detections, or similar appearances. Recent methods relying solely on motion (robust to short occlusions) or deep appearance embeddings (effective for long re-identification [2]) yield suboptimal performance in dynamic/crowded scenes. Transformer-based approaches [3, 4] enable long-range temporal modeling via global attention, but their $O(n^2)$ complexity limits real-time applicability, and they lack adaptation to multi-scale temporal structures of object dynamics.

This work is supported by the National Natural Science Foundation of China under Grant 62206046, 62477006, and 62473288.

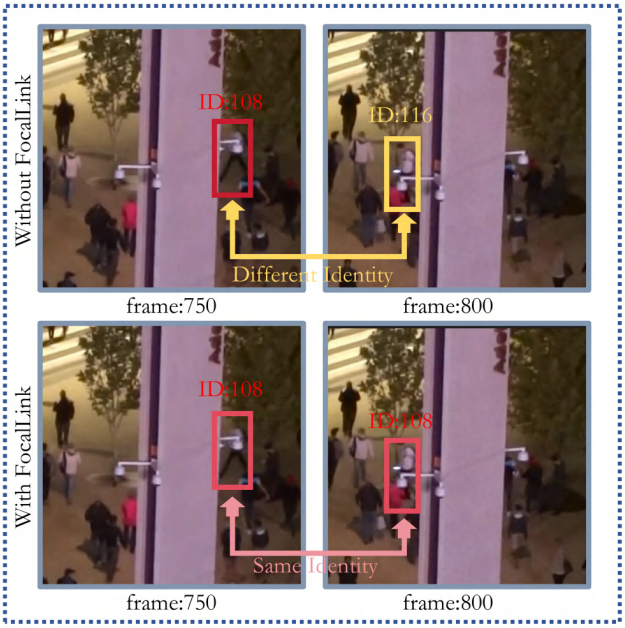

Fig. 1. During occlusion (frames 750–800), the tracker without FocalLink exhibits an identity switch from ID 108 to ID 116, whereas the tracker with FocalLink preserves ID 108, avoiding occlusion-induced identity switches.

To address the aforementioned issues, we propose FocalLink, a lightweight and plug-and-play trajectory association framework that can be seamlessly integrated into any existing tracker. The framework introduces a temporal focal modulation mechanism, adapted from spatial attention [5], which captures hierarchical temporal dependencies with linear complexity $O(n)$, achieving a balance between efficiency and scalability. In addition, it incorporates an identity-aware contrastive learning module that enhances the separability of embedding features across different objects. Extensive experiments on the MOT17 and MOT20 benchmarks demonstrate that FocalLink consistently improves the performance of state-of-the-art (SOTA) trackers without requiring substantial architectural modifications. It delivers higher tracking accuracy and significantly reduces identity switches, particularly in challenging scenarios with frequent occlusions and re-identification demands, thereby establishing new state-of-the-art results on the benchmarks.

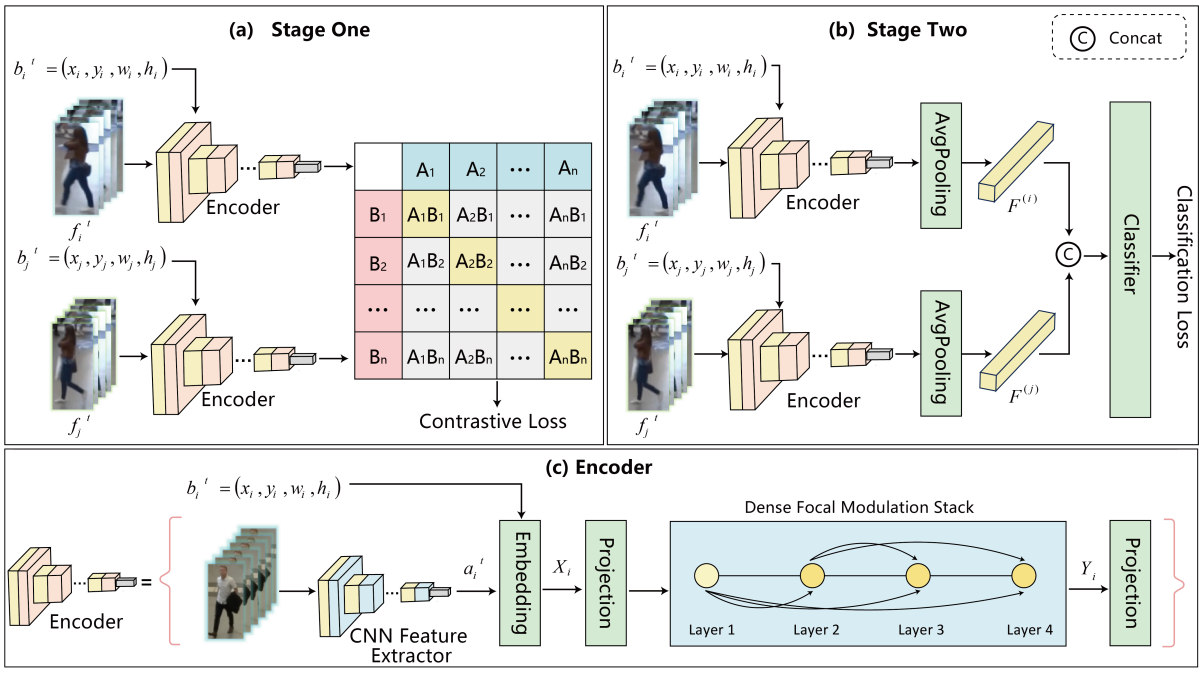

Fig. 2. Overview of FocalLink's two-stage training. (a) Stage One: the Dense Focal Modulation encoder is trained with contrastive loss while the classifier is frozen. (b) Stage Two: the full network is trained jointly, with contrastive loss weight gradually reduced to focus on association. (c) Encoder architecture: input features and box coordinates are extracted via a Convolutional Neural Network, projected, and passed through the Dense Focal Modulation Stack.

2. RELATED WORK

Most multi-object tracking (MOT) systems follow a tracking-by-detection paradigm, where objects are first detected in individual frames and then associated across time [1, 2]. Association strategies range from frame-level methods leveraging spatial cues, motion features, or ReID embeddings [6, 7, 8, 9] to trajectory-level approaches that merge fragmented tracklets over longer horizons [10, 11, 12]. While effective in moderate scenarios, these methods often struggle with long-term identity preservation under occlusions or missed detections. To address this, recent works have introduced Transformer-based architectures for global association, such as TrackFormer [4], TransTrack [3], TransCenter [13], and MOTR [14], as well as relational modeling methods like TransLink [15], GeneralTrack [16], and MOTIP [17]. However, their reliance on full self-attention incurs quadratic complexity, hindering real-time performance and scalability. Focal Modulation [5] offers a more efficient alternative by hierarchically aggregating local and global context with linear complexity, but its potential for temporal modeling remains underexplored. In this work, we extend focal modulation to the temporal domain with a dense modulation stack for tracklet features, enabling multi-scale dependency modeling that is particularly effective for handling prolonged occlusions, missing detections, and fragmented trajectories.

3. METHODOLOGY

3.1. Overall Architecture

FocalLink tackles trajectory association in multi-object tracking by combining focal modulation and contrastive learning to match fragmented trajectories under occlusions and interactions. As shown in Figure 2, the framework integrates temporal focal modulation for efficient multi-scale modeling, dense connections for feature reuse, and contrastive learning for discriminative identity embedding.

3.2. Problem Formulation

Let $\mathcal{T}_i = \{(f_i^t, b_i^t, a_i^t) | t = 1, 2, ..., T_i\}$ represent a trajectory segment, where f_i^t denotes the frame index, $b_i^t = (x_i, y_i, w_i, h_i)$ corresponds to the bounding box coordinates, and a_i^t represents appearance embeddings extracted via ReID networks. Given two trajectory segments $\mathcal{T}_i$ and $\mathcal{T}_j$, potentially from different time periods, determining whether they belong to the same physical entity constitutes the trajectory association problem.

3.3. Dense Focal Modulation Stack

Standard attention mechanisms are computationally expensive for temporal modeling due to their quadratic complexity with sequence length. To address this, we adopt the focal modulation mechanism [5], which achieves linear complexity while maintaining strong performance. As shown in Figure 2, it decomposes temporal modeling into hierarchical contextualization, gated aggregation, and element-wise integration. Building on this, we introduce the Dense Focal Modulation Stack, which employs densely connected modulation layers to enhance information flow and preserve identity features in trajectory modeling.

First, hierarchical contextualization captures multi-scale temporal dependencies through depth-wise separable convolutions with progressively expanding receptive fields:

$$\mathbf{C}^{(l)} = \mathrm{DWConv}^{(l)}(\mathbf{X}), \quad l = 1, \ldots, L, \tag{1}$$

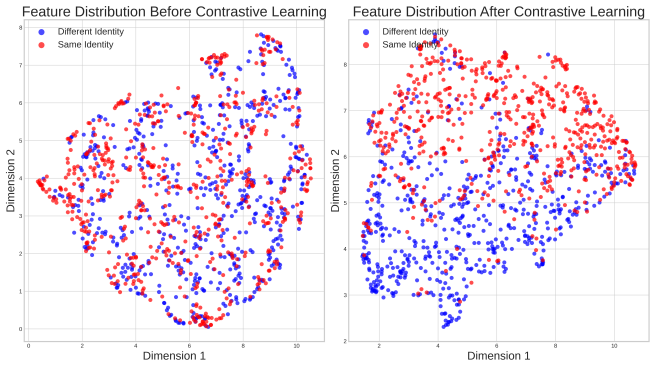

Fig. 3. UMAP visualization of embeddings before and after contrastive learning. Our approach leads to more compact same-identity clusters and better separation from other identities, highlighting improved feature discriminability.

where each level l operates with kernel size $K_l = K_1 \cdot s^{l-1}$ that grows by a factor s across levels. This creates a pyramid of temporal receptive fields that simultaneously captures fine-grained motion details and broader movement patterns. Subsequently, gated aggregation adaptively combines information across temporal scales:

$$\mathbf{C}_i = \sum_{l=1}^{L} \alpha_i^{(l)} \cdot \mathbf{C}_i^{(l)}, \quad \sum_{l=1}^{L} \alpha_i^{(l)} = 1, \tag{2}$$

with attention weights $\alpha_i^{(l)}$ computed through a learnable projection function with adaptive pooling and softmax normalization. This enables the model to focus on relevant temporal scales for each specific trajectory segment. Finally, element-wise modulation integrates the aggregated context with the original features:

$$\mathbf{Y}_i = \gamma \cdot \mathbf{X}_i + \beta \cdot \mathbf{C}_i, \tag{3}$$

where γ and β are learnable parameters that balance the preservation of identity-specific features with the incorporation of temporal context. As shown in figure 2, the Dense Focal Modulation Stack establishes direct connections from each layer to all subsequent layers:

$$\mathbf{H}_n = [\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{n-1}], \tag{4}$$

where $\mathbf{X}_0$ represents the initial trajectory features and $[\cdot]$ denotes feature concatenation along the channel dimension. Each layer applies focal modulation to its input:

$$\mathbf{X}_n = \text{FocalModulation}_n(\mathbf{H}_n). \tag{5}$$

This architecture preserves the complete feature hierarchy across network layers, ensuring both low-level motion cues and high-level temporal patterns contribute to association decisions. Improved gradient flow facilitates more effective training, and feature reuse preserves parameter efficiency even with the denser connectivity.

3.4. Contrastive Learning Framework

Reliable trajectory association requires discriminative feature representations that distinguish between different object identities while maintaining consistency within the same identity. To achieve this, we implement a contrastive learning framework that explicitly shapes the embedding space geometry. The framework optimizes a dual-objective function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{con}, \tag{6}$$

where the classification loss $\mathcal{L}_{cls}$ handles the binary association decision:

$$\mathcal{L}_{cls} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}), \tag{7}$$

and the contrastive component $\mathcal{L}_{con}$ enforces structural constraints on the embedding space:

$$\mathcal{L}_{con} = \frac{1}{B} \sum_{i=1}^{B} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \tag{8}$$

where z_i represents the L2-normalized embedding vector, $P(i)$ denotes the set of positive samples sharing the same identity as i, $A(i)$ encompasses all embeddings in the batch, and τ controls the distribution concentration. As illustrated in Fig. 3, UMAP visualization of embeddings before and after contrastive learning shows that our approach leads to more compact same-identity clusters and better separation from other identities, highlighting improved feature discriminability.

4. EXPERIMENTS

4.1. Datasets and Evaluation Metrics

We evaluate FocalLink on MOT17 and MOT20 [18] benchmarks. MOT20 focuses on crowded scenes with higher occlusion rates, presenting greater challenges for identity preservation. We employ five metrics: HOTA (unified detection and association quality), MOTA (overall tracking precision), IDF1 (identity consistency), ATA (association quality), and ID Switches (IDSW, trajectory fragmentation events).

4.2. Implementation Details

We implement FocalLink with 4 Focal Modulation blocks (focal window=3, focal level=1, focal factor=2). Our architecture combines 4D spatial features with 2048D ReID embeddings, projecting them into 512D space. Training follows two-stage approach: 200 epochs contrastive pre-training, then 300 epochs joint optimization using AdamW optimizer (lr=3×10^{-4}, batch size=32).

4.3. Main Results on MOT17 and MOT20

Table 1 summarizes the impact of integrating FocalLink into diverse MOT baselines on MOT17 and MOT20. Across all backbones, FocalLink consistently improves identity-related metrics, particularly IDF1 and ATA, while substantially reducing ID switches. On MOT17, gains include up to +0.88 in IDF1 and +2.39 in ATA, alongside reduced IDSW (e.g., FLWN: 962 $\rightarrow$ 850). On MOT20, improvements are more pronounced: TrackFormer achieves +2.20 IDF1, +9.74 ATA, and a reduction of 434 IDSW, while TransCenter improves by +2.18 IDF1, +5.19 ATA, and –552 IDSW. Notably, both classical association-based trackers (e.g., FLWN, FeatureSORT) and Transformer-based models (e.g., TrackFormer, TransCenter) benefit, highlighting FocalLink's plug-and-play compatibility and scalability. These results demonstrate that FocalLink is a universal enhancement module that significantly strengthens long-term identity preservation under occlusions and crowded scenarios, pushing the state of the art across heterogeneous tracking paradigms.

Table 1. Performance comparison between baseline trackers and our FocalLink-enhanced variants on MOT17 and MOT20 benchmarks. For each metric pair, we show the absolute difference in parentheses.

Tracker	HOTA↑	MOTA↑	IDF1↑	ATA↑	IDSW↓
Dataset MOT17					
FLWN [6]	77.255	91.433	85.753	64.400	962
FLWN+FocalLink	77.278 (+**0.023**)	91.417 (-**0.016**)	86.092 (+**0.339**)	66.793 (+**2.393**)	850 (-**112**)
PermaTrack [19]	60.759	74.641	70.412	46.284	1965
PermaTrack+FocalLink	60.838 (+**0.079**)	74.636 (-**0.005**)	70.451 (+**0.039**)	46.333 (+**0.009**)	1931 (-**34**)
TransCenter [13]	58.122	70.159	67.535	35.134	2049
TransCenter+FocalLink	58.660 (+**0.538**)	70.161 (+**0.002**)	68.411 (+**0.876**)	35.856 (+**0.722**)	2012 (-**37**)
Dataset MOT20					
FLWN [6]	77.506	94.050	89.221	70.810	1135
FLWN+FocalLink	77.633 (+**0.127**)	94.058 (+**0.008**)	89.620 (+**0.399**)	74.046 (+**3.236**)	1023 (-**112**)
FeatureSORT [7]	76.707	93.812	89.405	75.730	970
FeatureSORT+FocalLink	76.738 (+**0.031**)	93.819 (+**0.007**)	89.616 (+**0.211**)	77.083 (+**1.353**)	915 (-**55**)
FairMOT [20]	69.551	88.498	85.232	50.523	3898
FairMOT+FocalLink	69.634 (+**0.083**)	88.511 (+**0.013**)	85.690 (+**0.458**)	55.188 (+**4.665**)	3740 (-**158**)
TrackFormer [4]	62.715	81.094	73.391	52.352	1975
TrackFormer+FocalLink	63.122 (+**0.417**)	81.061 (-**0.033**)	75.586 (+**2.197**)	62.095 (+**9.743**)	1541 (-**434**)
TransCenter [13]	55.859	80.631	63.746	37.492	4310
TransCenter+FocalLink	57.027 (+**1.168**)	80.678 (+**0.047**)	65.921 (+**2.175**)	42.685 (+**5.193**)	3758 (-**552**)

Table 2. Ablation study on the MOT20 benchmark evaluating FocalLink with (w/) and without (w/o) contrastive learning (CL). Best results within each tracker group are highlighted in bold.

Baseline	CL	HOTA↑	MOTA↑	IDF1↑	ATA↑	IDSW↓
FLWN	w/o	77.417	94.055	89.211	73.584	1047
	w/	**77.633**	**94.058**	**89.620**	**74.046**	**1023**
FeatureSORT	w/o	75.764	93.624	85.448	69.818	1021
	w/	**76.738**	**93.819**	**89.616**	**77.083**	**915**
FairMOT	w/o	69.550	88.499	85.232	59.029	3881
	w/	**69.634**	**88.511**	**85.690**	**55.188**	**3740**
TransCenter	w/o	55.772	80.636	63.779	40.055	4125
	w/	**57.027**	**80.678**	**65.921**	**42.685**	**3758**

Table 3. Ablation study on the MOT20 benchmark evaluating FocalLink with (w/) and without (w/o) the Dense Focal Modulation Stack. Best results within each tracker group are highlighted in bold.

Baseline	Dense	HOTA↑	MOTA↑	IDF1↑	ATA↑	IDSW↓
FLWN	w/o	77.446	**94.061**	89.37	73.473	1027
	w/	**77.633**	94.058	**89.620**	**74.046**	**1023**
FeatureSORT	w/o	76.594	**93.820**	89.518	76.873	987
	w/	**76.738**	93.819	**89.616**	**77.083**	**915**
FairMOT	w/o	69.418	88.508	85.521	**55.916**	**3707**
	w/	**69.634**	**88.511**	**85.690**	55.188	**3740**
Trackformer	w/o	63.048	81.018	75.437	60.037	1590
	w/	**63.122**	**81.061**	**75.586**	**62.095**	**1541**

4.4. Ablation Studies

We conduct ablation studies examining: (1) the role of contrastive learning, and (2) the effect of dense connectivity.

4.4.1. Impact of Contrastive Learning

Table 2 evaluates the impact of the contrastive learning (CL) module within FocalLink on the MOT20 benchmark. Across all tested trackers, incorporating CL consistently improves tracking performance: HOTA and IDF1 increase, ATA is generally higher, and identity switches are reduced. For instance, FLWN gains +0.216 HOTA and reduces IDSW from 1047 to 1023, while FeatureSORT sees a +0.974 HOTA increase and 106 fewer ID switches. These results demonstrate that the contrastive learning component is crucial for enhancing identity discrimination and long-term trajectory association within FocalLink, confirming its effectiveness as an integral module rather than an optional addition.

4.4.2. Effect of Dense Connectivity

Table 3 presents the impact of the Dense Focal Modulation Stack within FocalLink on the MOT20 benchmark. Across all baseline trackers, incorporating the dense stack consistently improves overall tracking performance. HOTA and IDF1 scores increase for each tracker, and ATA is generally enhanced, while identity switches are reduced. For example, FLWN gains +0.187 HOTA and reduces IDSW from 1027 to 1023, and FeatureSORT sees HOTA increase from 76.594 to 76.738 with 72 fewer ID switches. These results demonstrate that the dense temporal aggregation provided by the stack effectively preserves identity information and strengthens multi-scale temporal modeling, confirming its essential role in FocalLink's trajectory association framework.

5. CONCLUSION

We propose FocalLink, a trajectory association framework that improves identity preservation via focal modulation, dense connectivity, and contrastive learning. Experiments on MOT17 and MOT20 show consistent performance gains across various trackers, including reduced identity switches and improved identity-aware metrics. FocalLink also enhances state-of-the-art MOT models, demonstrating its broad compatibility and general applicability. Future work will explore adaptive temporal modeling for long-term association.

6. REFERENCES

[1] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Tozeto Ramos, and Ben Upcroft, "Simple online and realtime tracking," *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

[2] 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)*, pp. 3645–3649, 2017.

[3] Pei Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo, "Transtrack: Multiple-object tracking with transformer," *ArXiv*, vol. abs/2012.15460, 2020.

[4] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer, "Trackformer: Multi-object tracking with transformers," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8834–8844, 2021.

[5] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao, "Focal modulation networks," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. 2022, vol. 35, pp. 4203–4217, Curran Associates, Inc.

[6] Hongbin Liu, Yongze Zhao, Peng Dong, Xiuyi Guo, and Yilin Wang, "Iof-tracker: A two-stage multiple targets tracking method using spatial-temporal fusion algorithm," *Applied Sciences*, vol. 15, no. 1, 2025.

[7] Hamidreza Hashempoor, Rosemary Koikara, and Yu Dong Hwang, "Featuresort: Essential features for effective tracking," 2024.

[8] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 1–21, Springer Nature Switzerland.

[9] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani, "Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification," *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3025–3029, 2023.

[10] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng, "Strongsort: Make deepsort great again," *IEEE Transactions on Multimedia*, vol. 25, pp. 8725–8737, 2022.

[11] Quanwei Yang, Xinchen Liu, Wu Liu, Hongtao Xie, Xiaoyan Gu, Lingyun Yu, and Yongdong Zhang, "Remot: A region-to-whole framework for realistic human motion transfer," in *Proceedings of the 30th ACM International Conference on Multimedia*, New York, NY, USA, 2022, MM '22, p. 1128–1137, Association for Computing Machinery.

[12] Gaoang Wang, Yizhou Wang, Renshu Gu, Weijie Hu, and Jenq-Neng Hwang, "Split and connect: A universal tracklet booster for multi-object tracking," *IEEE Transactions on Multimedia*, vol. 25, pp. 1256–1268, 2021.

[13] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda, "Transcenter: Transformers with dense representations for multiple-object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 7820–7835, 2021.

[14] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei, "Motr: End-to-end multiple-object tracking with transformer," in *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds., Cham, 2022, pp. 659–675, Springer Nature Switzerland.

[15] Yanting Zhang, Shuanghong Wang, Yuxuan Fan, Gaoang Wang, and Cairong Yan, "Translink: Transformer-based embedding for tracklets' global link," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[16] Zheng Qin, Le Wang, Sanpin Zhou, Panpan Fu, Gang Hua, and Wei Tang, "Towards generalizable multi-object tracking," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19004, 2024.

[17] Ruopeng Gao, Ji Qi, and Limin Wang, "Multiple object tracking as id prediction," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 27883–27893.

[18] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Qinfeng Shi, Daniel Cremers, Ian D. Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taix'e, "Mot20: A benchmark for multi object tracking in crowded scenes," *ArXiv*, vol. abs/2003.09003, 2020.

[19] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon, "Learning to track with object permanence," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10840–10849, 2021.

[20] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, pp. 3069 – 3087, 2020.