Supplementary Material for
Advancing Player Identification and Tracking with Global ID Fusion (GIF)

Karol Wojtulewicz* Minxing Liu* Niklas Carlsson
Linköping University, Linköping, Sweden
karol.wojtulewicz@liu.se, minxing.liu@liu.se, niklas.carlsson@liu.se

1. Additional Dataset Experiments

1.1. Oracle-Based Characterization

To disentangle the challenges in object localization and association, we conduct an oracle analysis by using ground-truth bounding boxes combined with different association strategies. This upper-bound evaluation helps identify the primary bottlenecks in multi-object tracking across datasets. Specifically, we compare three association methods: (1) IoU-based matching (computing IoU between bounding boxes in adjacent frames), (2) motion modeling via a Kalman Filter under linear motion assumptions, and (3) appearance matching using a pre-trained ReID model [3]. We evaluate these settings on MuPNIT and the closest dataset to it, SportsMOT [1].

The results are presented in Table 1. On SportsMOT, performance across all metrics is high, and notably, using only IoU matching yields the best results, indicating that most targets follow predictable and smooth motion patterns. This suggests that association is relatively straightforward in SportsMOT under oracle detection.

In contrast, MuPNIT shows markedly lower scores for association metrics (AssA, IDF1), despite near-perfect detection performance (DetA, MOTA) under oracle bounding boxes—underscoring that the challenge lies not in annotations (or detections), but in the real-world complexity of identity association under broadcast-style camera dynamics. Interestingly, adding naive appearance or motion cues fails to improve—and occasionally reduces—association performance. This can be attributed to the frequent visual similarity among subjects and complex, non-linear motion in MuPNIT sequences, revealing the limitations of traditional heuristics under MuPNIT's identity ambiguity and frequent camera switches. These results demonstrate that MuPNIT presents a significantly more difficult tracking scenario, requiring improved association techniques beyond conventional cues. Our GIF approach addresses these and other real-world challenges often not captured by prior datasets.

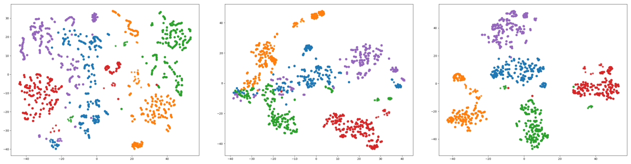

DeepSORT [6] CLIP$_{proj}$ [4] AP3D [2]

Figure 1. t-SNE [5] visualization of ReID embeddings for five randomly selected players in our MuPNIT-ReID-light dataset using three different appearance models: DeepSORT [6] (baseline), CLIP$_{proj}$ [4], and AP3D [2]. Circle points indicate non-zoom-in images, while star points denote zoom-in views. Better intra-ID clustering is observed with CLIP$_{proj}$ and AP3D. Best viewed in color.

1.2. MuPNIT-ReID Feature Visualization

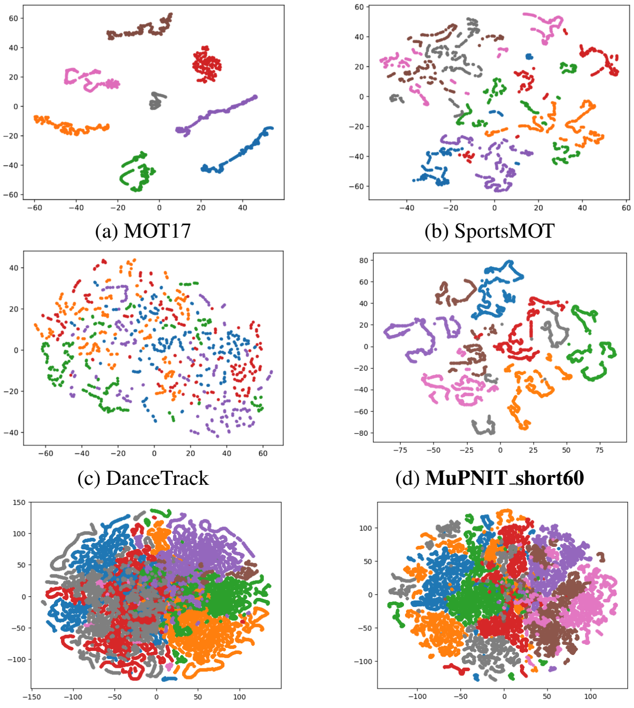
We randomly sample five identities from the MuPNIT-ReID-light dataset and visualize their ReID feature distributions using t-SNE in three different embedding spaces: DeepSORT (baseline), CLIP$_{proj}$, and AP3D (Figure 1). While DeepSORT exhibits mixed or loosely clustered identity representations, both CLIP$_{proj}$ and AP3D show more compact and well-separated clusters for the same identities. This highlights that stronger ReID encoders can help mitigate appearance variability in MuPNIT, but still require integration with robust tracking and ID-assignment strategies—as done in our GIF framework.

To further investigate the distinguishability of appearance features across different tracking datasets, we visualize the DeepSORT appearance embeddings of multiple identities using t-SNE [5], as shown in Figure 2. We include representative video sequences from MOT17, SportsMOT, DanceTrack, and three test sequences from MuPNIT. In MOT17, feature clusters are clearly separated, indicating that objects are visually distinguishable, making appearance-based association relatively easy. SportsMOT shows moderate entanglement between identities, while DanceTrack exhibits highly overlapped and intertwined clusters, consistent with its emphasis on uniform appearances and complex motion. In MuPNIT, embeddings show

*Equal Contribution

Table 1. Oracle analysis of different association models on SportsMOT and MuPNIT test set, respectively. **A**, **I** and **M** correspond to **A**ppearance, **I**oU and **M**otion respectively.

			SportsMOT					MuPNIT_60FPS test split					MuPNIT_30FPS test split				
A	**I**	**M**	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1	HOTA	DetA	AssA	MOTA	IDF1
	✓		80.203	97.59	65.913	96.014	73.354	41.544	**99.764**	17.3	99.679	24.781	41.502	**99.588**	17.295	99.358	24.766
	✓	✓	73.281	87.938	61.09	98.829	76.138	40.931	97.917	17.11	**99.845**	24.968	39.714	95.165	16.575	**99.689**	24.779
✓	✓	✓	70.076	82.319	59.671	97.451	77.107	39.119	93.25	16.414	99.674	24.838	37.998	89.93	16.062	99.313	24.937
✓			**87.577**	**99.583**	**77.018**	**99.566**	**82.98**	**41.826**	99.713	**17.544**	99.686	**25.193**	**42.031**	99.559	**17.744**	99.379	**25.692**



(a) MOT17

(b) SportsMOT

(c) DanceTrack

(d) **MuPNIT_short60**

(e) **MuPNIT_long60**

(f) **MuPNIT_long30**

Figure 2. t-SNE [5] visualization of DeepSORT appearance features from sampled videos in MOT17, DanceTrack, SportsMOT and MuPNIT datasets. Different IDs are coded by the different colors. The distinguishness of objects in MuPNIT lies between that of SportsMOT and DanceTrack.

moderate cluster separability—tighter than DanceTrack but more entangled than MOT17—highlighting its unique mix of visual similarity and temporal density, which challenges ReID more than most prior datasets. This aligns with our oracle analysis findings and confirms that although appearance cues offer some utility, they remain insufficient for reliable association in MuPNIT, especially in zoomed-in and fast-paced sports scenes. In addition, MuPNIT exhibits much denser data points compared to other datasets, highlighting the greater challenge of maintaining consistent identity tracking over longer temporal spans.

References

[1] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023. 1

[2] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. *ArXiv*, abs/2007.08434, 2020. 1

[3] Ziqiang Pei. Deepsort pytorch. https://github.com/ZQPei/deep_sort_pytorch, 2019. Accessed: 2025-07-14. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1, 2

[6] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. 1