DEPTH-AWARE SCORING AND HIERARCHICAL ALIGNMENT FOR MULTIPLE OBJECT TRACKING

Milad Khanchi, Maria Amer, and Charalambos Poullis

Supplementary Material (ICIP 2025)

1. BENCHMARK EVALUATION ON MOT17 AND MOT20

Main performance metrics for MOT are HOTA, IDF1, and AssA [1]. HOTA assesses both detection and association accuracy. IDF1 and AssA primarily evaluate association performance, while MOTA is predominantly focused on detection accuracy. In our results tables, ↑means the higher, the better, and ↓means the lower, the better. Bold numbers indicate the best performance. We adopt YOLOX [2] as the default object detector. In our results tables, MOT methods that also use YOLOX as the detector are highlighted in blue.

MOT17 and MOT20 are well-established pedestrian tracking benchmarks characterized by relatively linear motion patterns compared to DanceTrack and SportsMOT having non-linear motion. MOT17 consists of urban scenes with moderate crowd density and occasional occlusions, while MOT20 presents highly crowded environments. These datasets are widely used for evaluating tracking performance in dense, urban scenarios with largely predictable pedestrian trajectories. However, their focus on objects with temporally consistent orientations relative to the camera differs from the dynamic, non-linear motion patterns that DepthMOT is optimized for. We evaluate DepthMOT on MOT17 and MOT20 under the *private* detection protocol, as shown in Table 1.

MOTRv2 [3], a joint detection-ReID framework, designed for non-linear motion, underperforms on the linear motion datasets MOT17 and MOT20 (see Table 1). Trackers based on linear motion models, such as CMTrack [4], MotionTrack [5], and Deep OC-SORT [6], excel in MOT17 and MOT20 due to the predictable pedestrian trajectories. In contrast, DiffMOT [7] and our DepthMOT perform better in more dynamic scenarios like DanceTrack and SportsMOT (as shown in our main manuscript). Both non-linear models face challenges in MOT17 and MOT20, where linear motion models tend to be more effective.

Additionally, as illustrated in Fig. 1, zero-shot depth estimation encounters challenges in low-light conditions, producing low-contrast depth maps that make it difficult to distinguish distant objects. This limitation affects the accuracy of DepthMOT on MOT17 and MOT20. Despite these challenges, DepthMOT achieves the lowest false positive (FP) rate of 1.3, demonstrating the effectiveness of HAS in reducing false associations and track fragmentation.

While DepthMOT is not explicitly trained on each MOT dataset, it still performs competitively in high-occlusion environments and showcases strong adaptability across diverse tracking scenarios.

MOT17									
Tracker	**HOTA↑**	**MOTA↑**	**IDF1↑**	**FP(10^4)↓**	**FN(10^4)↓**	**ID$_s$↓**	**Frag↓**	**AssA↑**	**AssR↑**
FairMOT [8]	59.3	73.7	72.3	2.75	11.7	3,303	8,073	58.0	63.6
TransTrack [9]	54.1	75.2	63.5	5.02	8.64	3,603	4,872	47.9	57.1
MOTR [10]	57.2	71.9	68.4	2.11	13.6	2,115	3,897	55.8	59.2
CenterTrack [11]	52.2	67.8	64.7	1.8	1.6	3,039	-	-	-
MeMOTR [12]	58.8	72.8	71.5	-	-	-	-	58.4	-
DiffusionTrack [13]	60.8	77.9	73.8	-	-	3,819	4,815	58.8	-
MixSort-OC [14]	63.4	78.9	77.8	-	-	1,509	-	63.2	-
MixSort-Byte [14]	64.0	78.9	78.7	-	-	2,235	-	64.2	-
C-BIoU [15]	64.1	81.1	79.7	-	-	-	-	63.7	-
MOTRv2 [3]	62.0	78.6	75.0	-	-	-	-	60.6	-
GHOST [16]	62.8	78.7	77.1	-	-	2,325	-	-	-
ByteTrack [17]	63.1	80.3	77.3	2.55	8.37	2,196	2,277	62.0	68.2
OC-SORT [18]	63.2	78.0	77.5	1.51	10.8	1,950	2,040	63.2	67.5
StrongSORT [19]	63.5	78.3	78.5	-	-	1,446	-	63.7	-
GeneralTrack [20]	64.0	80.6	78.3	-	-	1,563	-	63.1	-
StrongSORT++ [19]	64.4	79.6	79.5	2.79	8.62	1,194	1,866	64.4	**71.0**
Deep OC-SORT [6]	64.9	79.4	80.6	1.66	9.88	1,023	2,196	65.9	70.1
MotionTrack [5]	65.1	**81.1**	80.1	2.38	**8.16**	1,140	-	-	-
CMTrack [4]	**65.5**	80.7	**81.5**	2.59	8.19	**912**	**1,653**	**66.1**	-
*DiffMOT [7]	64.5	79.8	79.3	-	-	-	-	64.6	-
*DepthMOT	62.7	76.5	77.9	**1.3**	11.7	1,342	-	63.6	68.8
MOT20									
Tracker	**HOTA↑**	**MOTA↑**	**IDF1↑**	**FP(10^4)↓**	**FN(10^4)↓**	**ID$_s$↓**	**Frag↓**	**AssA↑**	**AssR↑**
FairMOT [8]	54.6	61.8	67.3	10.3	8.89	5,243	7,874	54.7	60.7
DiffusionTrack [13]	55.3	72.8	66.3	-	-	4,117	4,446	51.3	-
MOTRv2 [3]	60.3	76.2	72.2	-	-	-	-	58.1	-
GHOST [16]	61.2	73.7	75.2	-	-	1,264	-	-	-
ByteTrack [17]	61.3	77.8	75.2	2.62	8.76	1,223	1,460	59.6	66.2
GeneralTrack [20]	61.4	77.2	74.0	-	-	1,627	-	59.5	-
StrongSORT [19]	61.5	72.2	75.9	-	-	1,066	-	63.2	-
OC-SORT [18]	62.1	75.5	75.9	1.80	10.8	913	1,198	62.0	67.5
StrongSORT++ [19]	62.6	73.8	77.0	1.66	11.8	770	1,003	64.0	69.6
MotionTrack [5]	62.8	78.0	76.5	2.86	**8.41**	1,165	1,321	61.8	-
Deep OC-SORT [6]	63.9	75.6	79.2	1.69	10.8	779	1,536	65.7	**70.8**
CMTrack [4]	**64.8**	76.2	**79.9**	2.22	10.04	**730**	**987**	**66.7**	-
*DiffMOT [7]	61.7	**76.7**	74.9	-	-	-	-	60.5	-
*DepthMOT	62.4	73.2	77.3	**1.3**	12	1,141	-	64.3	68.6

Table 1. **Results on MOT17-test and MOT20-test.** Methods in the blue blocks use the same YOLOX detector. The methods with * indicate that they are non-linear models. As can be seen, no tracker performs best across metrics and datasets. Our DepthMOT has the lowest false positive (FP).

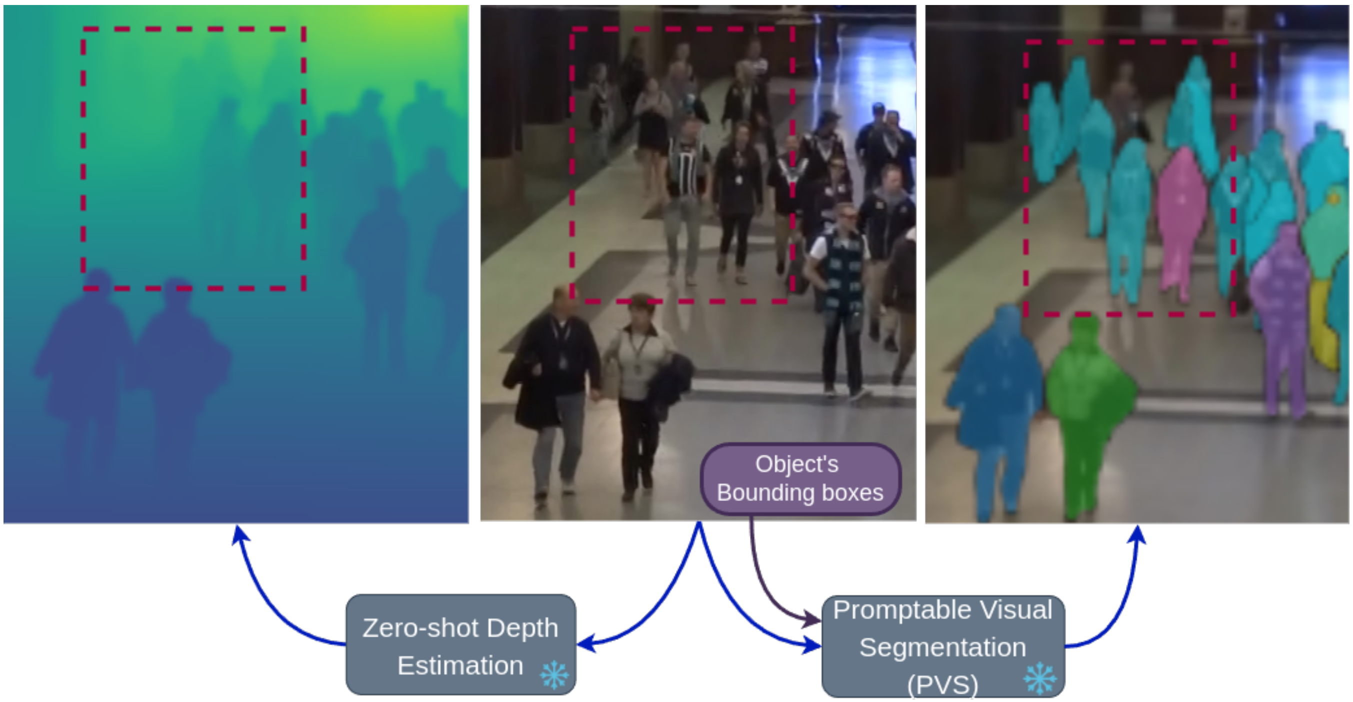

Fig. 1. **Challenges.** An example of zero-shot depth estimation and PVS modules, emphasizing the encountered challenges under different lighting conditions in the MOT20 dataset. The highlighted area, marked by a dotted square, illustrates that the depth map of certain objects is not accurately predicted.

2. VISUAL RESULTS

Fig. 2 and Fig.3 illustrate tracking performance of our DepthMOT framework on challenging sequences from the DanceTrack and SportsMOT datasets, respectively. Additionally, Fig. 4 and Fig.5 showcase the results of DepthMOT on MOT17 and MOT20. For enhanced visualization, bounding boxes of individuals with similar IDs are displayed in similar colors, with the unique ID indicated in red at the top of each bounding box. However, some colors may appear visually similar due to the high density of individuals in specific frames.

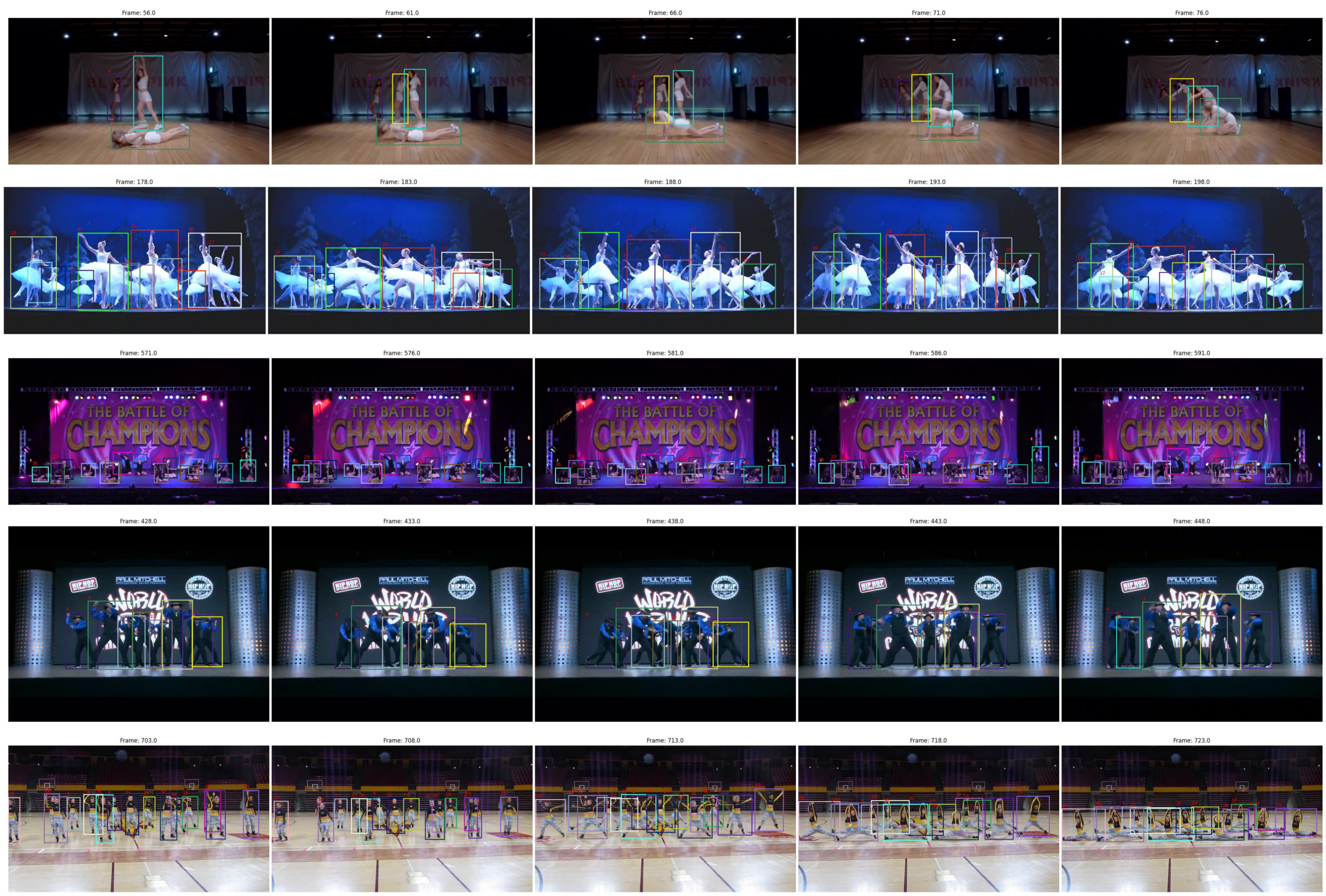

Fig. 2. **Examples.** DanceTrack tracking results

Fig. 3. **Examples.** SportsMOT tracking results

Fig. 4. **Examples.** MOT17 tracking results

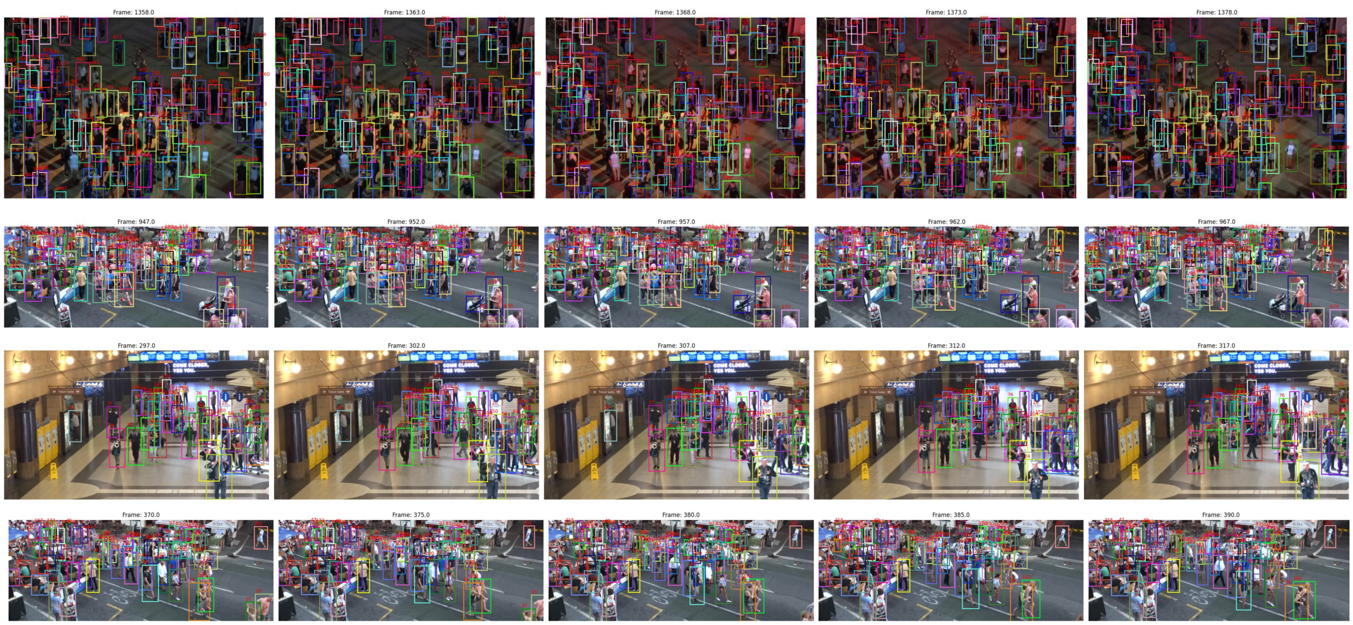

Fig. 5. **Examples.** MOT20 tracking results

3. REFERENCES

[1] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, 2021.

[2] G. Zheng, L. Songtao, W. Feng, L. Zeming, S. Jian *et al.*, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[3] Y. Zhang, T. Wang, and X. Zhang, "MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 056–22 065.

[4] K. Shim, J. Hwang, K. Ko, and C. Kim, "A Confidence-Aware Matching Strategy for Generalized Multi-Object Tracking," in *IEEE Int. Conf. Image Process.* IEEE, 2024, pp. 4042–4048.

[5] C. Xiao, Q. Cao, Y. Zhong, L. Lan, X. Zhang, Z. Luo, and D. Tao, "MotionTrack: Learning motion predictor for multiple object tracking," *Neural Networks*, vol. 179, p. 106539, 2024.

[6] G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification," in *IEEE Int. Conf. Image Process.* IEEE, 2023, pp. 3025–3029.

[7] W. Lv, Y. Huang, N. Zhang, R.-S. Lin, M. Han, and D. Zeng, "DiffMot: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 19 321–19 330.

[8] 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, pp. 3069–3087, 2021.

[9] P. Sun, J. Cao, Y. Jiang, R. Zhang, E. Xie, Z. Yuan, C. Wang, and P. Luo, "TransTrack: Multiple Object Tracking with Transformer," *arXiv preprint arXiv:2012.15460*, 2020.

[10] F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "MOTR: End-to-End Multiple-Object Tracking with Transformer," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 659–675.

[11] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking Objects as Points," in *Eur. Conf. Comput. Vis.* Springer, 2020, pp. 474–490.

[12] R. Gao and L. Wang, "MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking," in *Int. Conf. Comput. Vis.*, 2023, pp. 9901–9910.

[13] R. Luo, Z. Song, L. Ma, J. Wei, W. Yang, and M. Yang, "DiffusionTrack: Diffusion Model for Multi-Object Tracking," in *AAAI*, vol. 38, no. 5, 2024, pp. 3991–3999.

[14] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes," in *Int. Conf. Comput. Vis.*, 2023, pp. 9921–9931.

[15] F. Yang, S. Odashima, S. Masui, and S. Jiang, "Hard to Track Objects With Irregular Motions and Similar Appearances? Make It Easier by Buffering the Matching Space," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4799–4808.

[16] J. Seidenschwarz, G. Brasó, V. C. Serrano, I. Elezi, and L. Leal-Taixé, "Simple Cues Lead to a Strong Multi-Object Tracker," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 13 813–13 823.

[17] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object Tracking by Associating Every Detection Box," in *Eur. Conf. Comput. Vis.* Springer, 2022, pp. 1–21.

[18] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 9686–9696.

[19] Y. Du, Z. Zhao, Y. Song, Y. Zhao, F. Su, T. Gong, and H. Meng, "StrongSORT: Make DeepSORT Great Again," *IEEE Trans. Multimedia*, vol. 25, pp. 8725–8737, 2023.

[20] Z. Qin, L. Wang, S. Zhou, P. Fu, G. Hua, and W. Tang, "Towards Generalizable Multi-Object Tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 18 995–19 004.