GMT: Effective Global Framework for Multi-Target Multi-Camera Tracking

Supplementary Material

A. Expanded Related Work

A.1. MCMT Tracking in 3D Space

Some methods perform 3D reconstruction using inputs from multiple views and track targets in 3D space, achieving a similar effect of multi-camera multi-target tracking. Yang et al. [9] proposed a hybrid trajectory generation strategy for multi-camera tracking in gymnastics scenarios. The model performs traditional triangulation when sufficient detections are available, while prior knowledge of gymnastic motion is incorporated to assist trajectory generation under sparse detection conditions, effectively avoiding 3D reconstruction errors caused by limited detections. MITracker [7] adaptively fuses features from different views based on their quality and visibility, integrating them into the BEV space. The fused features are further enhanced through spatial-enhanced attention, enabling the model to focus on important regions of the target. MVTrajecter [8] enhances temporal modeling of both appearance and motion by enabling interaction between historical trajectories and current targets through the TMC and TAC modules, thereby improving its performance under occlusion and crowded scenarios.

Our proposed GMT focuses on MCMT tracking in 2D space. Although achieves a similar effect, GMT (or 2D-based methods) still differs significantly from the aforementioned 3D-based methods. Specifically, 3D-based methods rely on more information (i.e., camera intrinsic and extrinsic parameters, as well as 3D annotations for supervision), which generally makes them more accurate. However, this reliance on additional information also limits their practical applicability (e.g., obtaining camera intrinsics can be challenging when using moving cameras). Moreover, 3D-based methods typically demand significantly more computational resources. 2D-based methods are more flexible since they generally rely only on visual representations. They can operate even when timestamps across different devices are not strictly synchronized, a scenario in which reliable 3D reconstruction would be difficult to achieve.

In addition, due to differences in datasets, 3D-based and 2D-based methods cannot be directly compared. Specifically, 3D-based trackers require 3D annotations for training, but existing 2D datasets do not provide such labels, making it impossible for 3D-based trackers to run on 2D datasets. Conversely, 3D datasets typically evaluate cross-view tracking performance through top-down view tracking [5] which 2D-based trackers are unable to perform.

B. VisionTrack Dataset

B.1. Ethical Statement

All identifiable individuals in the dataset are from the institutions involved in this paper, and they are fully informed about the data collection process and its intended use.

B.2. Visualization

It is worth noting that although previous datasets also include multiple scenes, these scenes are usually limited to different areas within the same facility, resulting in relatively homogeneous backgrounds (as shown in Fig. A). In contrast, VisionTrack truly enables data collection across diverse environments, offering significantly greater diversity.

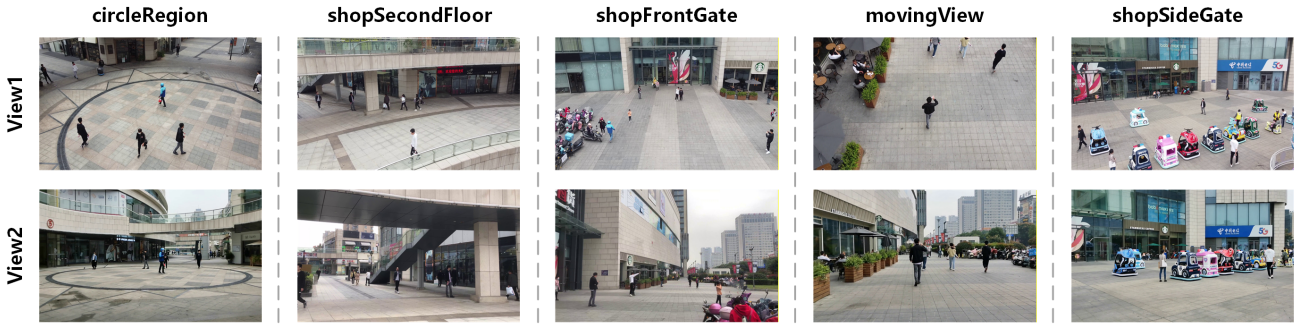
B.3. Limitations and Future Work

Although VisionTrack offers significant advantages in terms of diversity and scale, the use of drones for data collection increases scheduling complexity and cost, resulting in a smaller number of camera views compared with other datasets. Based on the experience of building VisionTrack, we are now working on constructing a new dataset with **more camera views** (including **multiple ground cameras** and **multiple UAV cameras** at varying altitudes), greater camera motion, enhanced diversity, and increased scale.

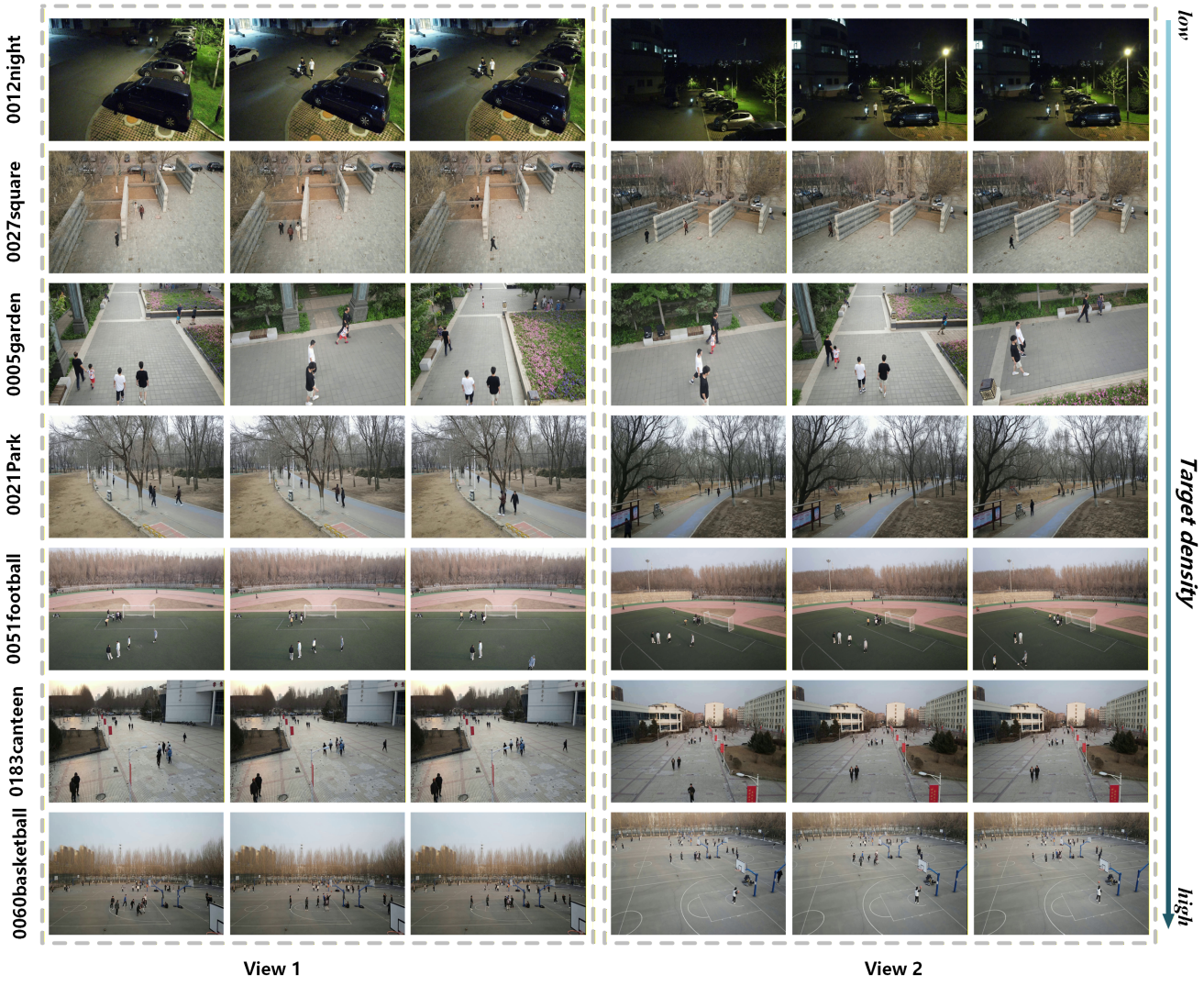
C. More Details about the GTA Module

C.1. ID Embedding

Compared with DETR-based single-view trackers that only use information from the previous frame, GMT maintains global trajectories that incorporate targets from multiple historical frames and multiple views, which naturally motivates us to explore strategies to reduce interference among targets and enhance trajectory modeling. To achieve this, we assign the same learnable trajectory ID embedding to all targets belonging to the same trajectory, allowing the model to recognize which trajectory each target belongs to. To prevent ID embeddings from learning information about specific trajectories during training, we create a set of embeddings much larger than the number of targets and randomly assign embeddings to trajectories in each iteration. As shown in Tab. B, introducing the trajectory embeddings **(w/o ID emb)** improves both multi-view matching and single-view tracking, demonstrating their effectiveness in improving trajectory modeling. We also observe that introducing ID embeddings brings more significant improvements in scenes with fewer targets, while their effect is lim-



(a) Visualization of different scenes in DIVOTrack



(b) Visualization of different scenes in VisionTrack

Figure A. **Visualized comparison between VisionTrack and DIVOTrack.** To address the issue of (a) strong homogeneity across different scenes in current datasets, we ensured that (b) the scenes in VisionTrack are as diverse as possible.

Parameter		CA↑	C1↑	Parameter		CA↑	C1↑	Parameter		CA↑	C1↑
Feature Dim	0	74.0	79.8	Traj Length	15	74.5	80.5	Learning rate	5×10^{-3}	70.2	74.7
	64	74.5	80.6		20	74.7	81.3		10^{-3}	73.9	80.2
	128	**75.2**	**81.3**		32	**75.2**	81.3		10^{-4}	**75.2**	**81.3**
	256	74.7	81.0		40	74.6	**81.5**		10^{-5}	74.6	80.4
Parameter		**CA↑**	**C1↑**	**Parameter**	$N_{\text{enc}} : N_{\text{dec}}$	**CA↑**	**C1↑**	**Parameter**		**CA↑**	**C1↑**
Num of Heads	1	73.4	80.4	Layers	1:1	**75.2**	**81.3**	Batch size	2	69.8	76.4
	4	74.1	80.3		2:1	74.2	80.1		4	72.5	79.6
	8	**75.2**	**81.3**		1:2	74.3	80.1		6	74.4	80.2
	16	74.2	81.3		2:2	74.8	81.0		8	**75.2**	**81.3**

Table A. Experiments of hyperparameters on VisionTrack. The best results are shown in **bold**.

Method	CVMA	CVIDF1	MOTA	HOTA	IDF1
w/o ID emb	74.6	80.6	77.5	65.5	81.3
+ intra attn	**75.4**	80.9	**78.1**	**66.5**	81.6
intra attn	74.3	79.2	77.7	65.2	80.7
camel	71.8	76.6	77.3	62.2	77.5
GMT	75.2	**81.3**	78.0	66.2	**82.1**

Table B. Ablation results of the CFVE module on VisionTrack.

ited in crowded scenarios with more targets. Further analysis will be provided in Sec. C.2.

C.2. Attempts to Enhance Trajectory Modeling

As shown in Tab. B, besides using ID embeddings, we explored several other approaches to enhance trajectory modeling: 1) applying an additional encoder to each trajectory individually after the main encoder that processes all trajectories together **(+ intra attn)**; 2) encoding each trajectory independently without inter-trajectory interaction **(intra attn)**; and 3) adopting the trajectory modeling strategy from CAMELTrack [4] **(camel)**. However, adding extra intra-trajectory encoding had almost no effect on the final results, while using only intra-trajectory encoding or more complex trajectory modeling methods even led to performance degradation. We also noticed that these approaches have a more positive effect on sequences with fewer targets than on those with many targets.

We believe that the above phenomenon is caused by interference from incorrect associations. Through observation, we found that in most cases, errors in existing single-view models mainly caused by detection rather than association (as also addressed in MOTRv2[11]), so association errors have a limited impact on trajectory modeling in single-view tracking. However, MCMT tracking introduces additional cross-view association tasks, which makes it more prone to failures (i.e. associate different targets as the same trajectory or splitting the same target into multiple trajectories). Specifically, we leverage all trajectories from the past T frames in GMT. While this design helps handle short-term occlusions and incorrect ID switches, it also increases the likelihood of introducing erroneous trajectories. When the trajectory length is set to 30 frames, scenes in VisionTrack with about 40 targets across two views can yield over 200 trajectories. In WildTrack, which has seven views with roughly 40 targets, setting the trajectory length to 20 frames could contain more than 300 trajectories. Based on this, we believe that since all trajectories are encoded simultaneously using the same encoder in the GTA module, the model can adaptively learn the differences between intra-trajectory features and inter-trajectory features through the attention mechanism. On the other hand, overly strong intra-trajectory modeling may amplify the interference from incorrect associations. This assumption is also consistent with our observation that intra-trajectory modeling shows different effects in sequences with varying numbers of targets. In the future, we will focus on exploring methods to eliminate errors within trajectories and further enhance intra-trajectory modeling based on our proposed global tracking framework.

D. More Experiments

D.1. Experiments on Hyperparameters

As shown in Tab. L, we conduct experiments on the following hyperparameters: 1) the dimension of relative position features **(Feature dim)**; 2) the trajectory length during inference **(Traj length)**; 3) the learning rate **(Learning rate)**; 4) the number of attention heads in the GTA model **(Num of heads)**; 5) the number of encoder and decoder layers in the GTA model **(Layers)**; 6) the trajectory length during training **(Batchsize)**. We use 128 as the dimension of the relative position features and set the maximum trajectory length during inference to 30. It is worth noting that since we apply the cosine learning rate schedule, GMT is relatively insensitive to the learning rate and could achieve the

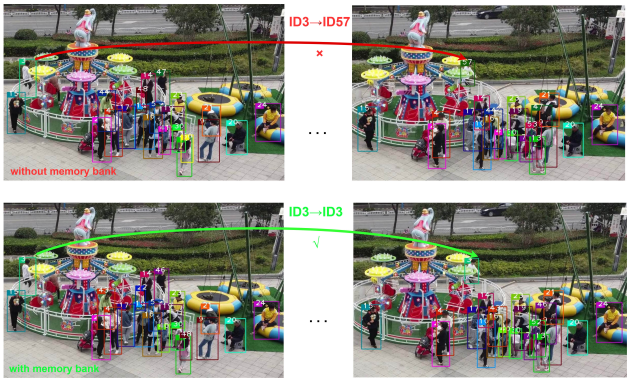


Figure B. Effect of the memory module.

best performance within the range of 5×10^{-4} to 5×10^{-5}.

D.2. Detailed Comparisons With SOTA Methods

Since existing MCMT trackers are usually evaluated on a limited number of datasets and often do not report all metrics, we conduct a more comprehensive comparison by reproducing their results across all major MCMT tracking benchmarks and presenting complete evaluation metrics. We hope our experiments can serve as a valuable reference for future research. We understand that readers may have concerns that the chosen baseline methods may seem relatively dated. However, as research on MCMT tracking remains limited, we have included nearly all recent and representative studies available. We believe that our simple yet effective global tracking framework along with the newly introduced dataset, can inspire further progress in the field of MCMT tracking.

VisionTrack. Compared with existing datasets, VisionTrack includes sequences captured across diverse scenarios, ranging from simple scenes with few non-overlapping targets to dense crowded scenes. This makes it more suitable for training and enables more comprehensive evaluation. VisionTrack adopts a novel setup with two moving drones capturing the scene from a top-down perspective. However, since the cameras share a large overlapping area and the top-down views provide a more complete observation of the scene, the cross-view association task in VisionTrack may be less challenging than in some other datasets. As shown in Tab. C, our proposed GMT achieves consistent performance improvements over other methods.

DIVOTrack [3]. As shown in Tab. D, GMT also achieves significant performance improvements. DIVOTrack consists of two handheld camera views and one drone view. Compared to other datasets, it contains relatively sparse target density within each view and contains almost no occlusion. Meanwhile, the large viewpoint differences make it more suitable for evaluating cross-view association rather than single-view tracking.

WildTrack [1]. Similar to VisionTrack, WildTrack also contains multiple views with high overlap. However, it features extremely high target density, severe occlusion, and a significantly larger number of views compared to other datasets, making it the most challenging among existing datasets. As shown in Tab. E, other trackers exhibit substantial performance degradation on WildTrack, while GMT maintains relatively balanced performance, demonstrating that the proposed global tracking framework better aligns with the nature of multi-camera tracking.

MvMHAT [5]. MvMHAT consists of seven sequences captured within the same scene. Although it is the simplest dataset compared to others, the targets share highly similar appearances, making it still somewhat challenging. As shown in Tab. F, GMT achieves the best performance on this dataset. It is worth noting that MvMHAT does not annotate severely occluded targets, yet we found that the detector can still detect some of them, which may affect the final evaluation metrics.

CAMPUS & EPFL [2]. Compared with other datasets, EPFL and CAMPUS have the largest inter-view differences, smaller overlapping regions, and relatively severe occlusions. As shown in Tabs. G and H, GMT achieves the best performance on most metrics, especially in terms of ID consistency and cross-view matching. Although GMT does not achieve the best results on some detection-related metrics (i.e. MOTA, MOTP, and DETA) or on some simpler metrics (i.e. IDsw, MT, ML), its performance remains highly competitive.

D.3. Experiment about the Memory Module

Since the memory module is only activated in very few cases compared with the entire dataset, and current multi-object tracking methods assign ground-truth labels based on IoU, which can easily fail under occlusion, it is difficult to quantify the effect of the memory module through evaluation metrics. Instead, we illustrate its impact through visualization. As shown in Fig. B, the memory module helps recover targets that have been occluded for a long time.

D.4. Experiment about Distance Threshold in RPCE

Tab. I shows the impact of the distance threshold range in RPCE on tracking performance. The **thres** represents the ratio between the threshold's lower and upper bounds and the square root of the target box area. It can be observed that when the distance threshold is too large (resulting in a wider sampling range), the tracking performance drops noticeably. This demonstrates that the threshold mechanism effectively filters out interference from targets located outside the overlapping region (i.e., farther away).

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML↓
AGW [10]	**70.1**	68.4	70.0	**80.3**	59.5	62.9	57.0	**77.0**	3482	526	54
CT [6]	65.0	62.7	76.7	78.8	55.8	62.2	50.9	65.5	3996	**544**	**37**
MvMHAT$_{prev}$ [2]	39.2	53.6	76.9	**80.4**	56.3	62.5	51.3	66.0	5136	523	54
CrossMOT [3]	64.5	64.4	75.6	78.5	54.5	**68.0**	49.3	65.8	3153	531	49
MvMHAT$_{Res}$[5]	51.1	61.2	77.0	79.9	56.0	62.3	50.8	66.4	4002	524	51
MvMHAT$_{Trans}$[5]	68.8	**70.0**	**77.6**	80.0	**59.6**	62.7	**57.2**	65.9	**2761**	527	52
GMT	**75.2**	**81.3**	**78.0**	79.6	**66.2**	**63.3**	**69.4**	**82.1**	**3192**	**550**	**37**

Table C. Comparison with state-of-the-art MCMT trackers on **VisionTrack**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML
AGW [10]	60.8	63.9	75.9	80.9	57.6	62.8	53.1	68.5	1740	374	69
CT [6]	67.6	66.5	75.6	80.7	58.2	62.6	54.4	70.4	1889	378	68
MvMHAT$_{prev}$ [2]	64.3	57.7	78.4	**82.7**	**60.8**	**66.3**	55.0	70.5	2440	460	53
CrossMOT [3]	69.0	69.1	77.8	81.5	60.5	64.2	**57.7**	73.9	**1431**	374	69
MvMHAT$_{Res}$[5]	67.5	65.3	78.9	81.4	58.1	65.1	51.8	70.1	1870	467	**53**
MvMHAT$_{Trans}$[5]	**69.4**	**68.6**	**78.9**	81.4	60.6	65.4	56.3	74.1	1853	**467**	53
GMT	**74.5**	**73.2**	**80.5**	**82.0**	**64.5**	**67.1**	**62.1**	**76.7**	**1565**	**470**	**52**

Table D. Comparison with state-of-the-art MCMT trackers on **DIVOTrack**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML
AGW [10]	15.9	23.2	52.6	78.8	49.8	48.5	**57.9**	63.7	**242**	64	249
CT [6]	17.3	27.7	53.6	**81.4**	**51.6**	46.3	57.5	**64.3**	**226**	67	**126**
MvMHAT$_{prev}$ [2]	9.4	15.9	47.4	80.7	43.4	43.0	44.1	51.6	422	55	249
CrossMOT [3]	**40.4**	**54.8**	**55.6**	78.8	49.8	**48.5**	52.7	63.4	369	**77**	142
MvMHAT$_{Res}$[5]	-	-	-	-	-	-	-	-	-	-	-
MvMHAT$_{Trans}$[5]	-	-	-	-	-	-	-	-	-	-	-
GMT	**61.7**	**72.0**	**64.9**	**91.2**	**56.1**	**53.3**	**59.1**	**74.1**	491	**131**	**135**

Table E. Comparison with state-of-the-art MCMT trackers on **WildTrack**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML
AGW [10]	89.6	**90.4**	89.7	90.4	71.0	83.0	60.9	73.1	151	**171**	0
CT [6]	48.9	56.6	93.7	89.5	74.9	84.2	66.7	80.9	483	171	0
MvMHAT$_{prev}$ [2]	82.4	71.8	93.1	**90.7**	75.7	**85.0**	67.5	82.8	170	170	0
CrossMOT [3]	90.9	86.1	**95.9**	86.6	78.0	82.3	**74.0**	**90.5**	123	171	0
MvMHAT$_{Res}$[5]	89.0	85.9	92.0	89.2	82.0	83.4	70.6	86.9	**109**	108	0
MvMHAT$_{Trans}$[5]	**89.6**	88.9	91.8	89.1	**82.3**	83.2	71.6	87.7	144	108	**0**
GMT	**94.1**	**95.6**	**96.2**	**93.1**	**87.5**	**89.8**	**85.3**	**95.8**	**82**	**171**	**0**

Table F. Comparison with state-of-the-art MCMT trackers on **MvMHAT**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML
AGW [10]	51.2	50.5	58.5	77.3	41.9	53.1	33.4	50.2	1710	61	**9**
CT [6]	58.5	47.8	59.0	77.6	42.7	53.1	34.6	52.3	1702	59	**9**
MvMHAT$_{prev}$ [2]	45.7	43.6	60.0	77.6	40.6	52.9	31.4	49.7	1938	68	13
CrossMOT [3]	**65.4**	**52.4**	**72.2**	**78.3**	**46.9**	**59.2**	35.8	**56.6**	881	**90**	13
MvMHAT$_{Res}$[5]	50.8	45.7	66.3	76.0	45.9	54.6	34.7	48.9	**720**	66	55
MvMHAT$_{Trans}$[5]	58.2	51.2	68.5	76.1	44.7	54.5	**37.1**	55.8	**733**	65	34
GMT	**66.4**	**66.8**	**69.5**	**77.8**	**52.3**	**56.7**	**48.2**	**69.1**	824	**79**	12

Table G. Comparison with state-of-the-art MCMT trackers on **CAMPUS**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	MOTP	HOTA	DETA	ASSA	IDF1	IDsw↓	MT	ML
AGW [10]	51.6	39.9	70.7	76.9	30.6	59.1	16.1	32.5	5948	89	**2**
CT [6]	40.5	28.6	69.3	76.1	30.5	58.1	16.2	33.4	6538	81	5
MvMHAT$_{prev}$ [2]	29.1	20.3	63.4	77.2	22.5	56.3	9.1	21.3	11522	70	**3**
CrossMOT [3]	**70.8**	**59.1**	77.6	**77.3**	**36.2**	**62.0**	**21.3**	**38.2**	3879	101	3
MvMHAT$_{Res}$[5]	31.1	23.5	**78.5**	77.1	30.1	62.0	14.8	31.6	**2760**	**109**	3
MvMHAT$_{Trans}$[5]	65.9	35.0	77.6	77.0	34.6	61.7	19.8	37.7	**1101**	108	3
GMT	**73.9**	**66.7**	**78.2**	**77.7**	**51.2**	**62.1**	**42.2**	**67.5**	4234	**108**	3

Table H. Comparison with state-of-the-art MCMT trackers on **EPFL**. The best two results are shown in **red** and **blue**.

Method	CVMA	CVIDF1	MOTA	HOTA	IDF1
1:3	74.9	80.8	77.6	65.7	81.6
2:4	**75.2**	**81.3**	**78.0**	**66.2**	**82.1**
2:6	73.6	77.5	79.5	64.3	79.6
4:6	72.1	76.2	77.7	63.0	78.5

Table I. Ablation results of the distance threshold on VisionTrack.

D.5. Experiment about the Trajectory Length

Compared with existing methods, our global trajectories incorporate target features from the past 30 frames, providing richer temporal context. Although this advantage is one of the primary motivations behind our design, we still conduct experiments under shorter trajectory lengths for a more comprehensive evaluation of GMT. As shown in Tab. J, even when using only the minimal temporal context, GMT consistently achieves the best tracking performance, surpassing all existing methods reported in Tab. C. These results verify the effectiveness of GMT's other designs, such as use of global information and relative position cues.

thres	CVMA	CVIDF1	MOTA	HOTA	IDF1
1	72.3	77.6	77.2	61.5	78.0
2	72.4	77.5	77.4	61.5	78.3
5	72.8	78.2	79.5	63.0	79.6
10	73.6	79.2	77.7	64.3	80.9

Table J. Ablation results of the trajectory length on VisionTrack.

D.6. Results of Joint Training

GMT adopts a two-stage training strategy, which mainly accelerates the speed of convergence through two mechanisms: (1) Since the detector is trained in the first stage, it can already provide accurate inputs for the RPCE and GTA modules at the beginning of the second stage; (2) The two-stage training avoids potential conflicts between detection and association losses.

We also experimented with a simpler end-to-end alternative in which detection and tracking are jointly optimized. In this joint training setting, we do not supervise VFCE and RPCE features separately. Instead, we directly apply the $loss_{VFCE}$ to the concatenated association features from both branches. As shown in the table, with 20k iterations (same as the main paper), the model exhibits only a slight performance drop, and at 30k iterations its performance becomes almost the same with that of the two-stage training.

	CVMA	CVIDF1	MOTA	HOTA	IDF1
20k	73.1	78.6	77.5	63.2	80.0
30k	74.8	81.0	78.2	66.2	81.7

Table K. Effect of Joint Training on VisionTrack.

D.7. Parameter and Speed Evaluation

Table L presents the parameters and flops of each module. It can be observed that GMT achieves a substantial performance improvement with only a marginal increase in computational cost over the detector, demonstrating the superiority of our proposed global framework. It should be noted that existing models often include many heuristic strategies that are not counted as parameters, which makes their parameter count appear smaller. Additionally, since these models typically rely on more complex loss functions to constrain cross-view association, they tend to be more difficult to converge than our global framework.

	Detector	VFCE	RPCE	GTA
Flops	227.1G	0.2G	0.8M	0.1G
Params	24.3M	0.7M	0.03M	5M

Table L. The number of parameters and Flops of each module.

D.8. Visualization

To more intuitively demonstrate the effectiveness of GMT, we provide additional visual results. Fig. C correspond to the visualization examples presented in Section **6** of the main paper. Fig. D shows the performance of GMT on other challenging sequences. It can be observed that even under conditions of overlapping small targets and large viewpoint variations, GMT still produces almost no errors, fully demonstrating the superiority of the global tracking framework.

References

[1] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *CVPR*, pages 5030–5039, 2018. 4

[2] Yiyang Gan, Ruize Han, Liqiang Yin, Wei Feng, and Song Wang. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM international conference on multimedia*, pages 282–290, 2021. 4, 5

[3] Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *IJCV*, 132(4):1075–1090, 2024. 4, 5

[4] Vladimir Somers, Baptiste Standaert, Victor Joos, Alexandre Alahi, and Christophe De Vleeschouwer. Cameltrack: Context-aware multi-cue exploitation for online multi-object tracking. *ArXiv*, abs/2505.01257, 2025. 3

[5] Ruize Han Zekun Qian Song Wang Wei Feng, Feifan Wang. Unveiling the power of self-supervision for multi-view multi-human association and tracking. In *arXiv*, 2023. 1, 4, 5

Figure C. The original visualization results of Figure 5 in the main text.

Figure D. More visualization results on VisionTrack.

[6] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV*, page 212–223, Berlin, Heidelberg, 2021. Springer-Verlag. 5

[7] Mengjie Xu, Yitao Zhu, Haotian Jiang, Jiaming Li, Zhenrong Shen, Sheng Wang, Haolin Huang, Xinyu Wang, Qing Yang, Han Zhang, and Qian Wang. Mitracker: Multi-view integration for visual object tracking. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27176–27185, 2025. 1

[8] Taiga Yamane, Ryo Masumura, Satoshi Suzuki, and Shota Orihashi. Mvtrajecter: Multi-view pedestrian tracking with trajectory motion cost and trajectory appearance cost. *International Conference on Computer Vision (ICCV)*, abs/2509.01157, 2025. 1

[9] Fan Yang, Shigeyuki Odashima, Shoichi Masui, Ikuo Kusajima, Sosuke Yamao, and Shan Jiang. Enhancing multi-camera gymnast tracking through domain knowledge integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):13386–13400, 2024. 1

[10] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE PAMI*, 44(6): 2872–2893, 2021. 5

[11] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22056–22065, 2023. 3