ReST: A Reconfigurable Spatial-Temporal Graph Model for Multi-Camera Multi-Object Tracking
—– Supplementary Material —–

In this supplementary material, we further describe additional details to complement our proposed reconfigurable graph model. Firstly, the calculation of the projection function is detailed in Section A, followed by tracklet ID assignment steps in Section B. Afterward, our network architecture and model complexity are described in Section C. Section D shows the effectiveness of post-processing module. Additionally, we provide visualization of our graph model to better realize two association stages in Section E. Qualitative results are shown in Section F to explain how we fix the fragmented tracklet problems. Lastly, we demonstrate the proposed ReST tracker in the attached link.

A. Calculation of Geometry Position

The geometry position of node v_i can be obtained by projecting its estimated foot point from the camera view to a common ground plane via a projection function. The projection function $P_{c_{v_i}}$ is based on camera calibration parameters and derived from perspective projection:

$$x_{img} = K[R \quad t]x_{world} = Px_{world}, \tag{1}$$

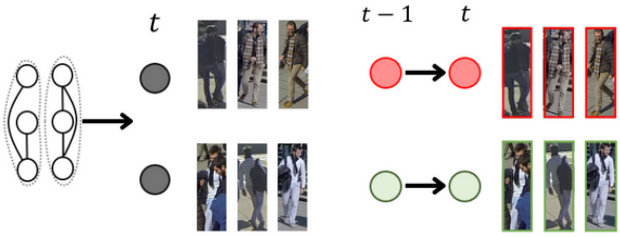
where x_{img} and x_{world} denote the positions in 2D image and 3D world represented in homogeneous coordinates, respectively, and K is the intrinsic matrix, with rotation matrix R and translation vector t determined by the extrinsic parameters. Assuming $z = 0$ as the common ground plane, the 3×4 projection matrix P is reduced to be a 3×3 homography matrix H. Therefore, the position p_{v_i} on the ground plane can be calculated by

$$p_{v_i} = H^{-1}x_{img}, \tag{2}$$

where x_{img} is the person's foot point, estimated by the position, width, and height of the bounding box.

B. Tracklet ID Assignment

In this section, we explain the details of tracklet ID assignment steps. In the spatial graph, one node v_i represents one detection, including bounding-box location b_{v_i} and camera ID c_{v_i}. In Graph Reconfiguration stage, we save all bounding-boxes and their respective camera IDs within the same connected component of spatial graph G_t^S, and then aggregate them into a node of G_t^T (Figure 1a). The last step of post-processing in Temporal Association is assigning tracklet ID. As shown in Figure 1b, if an aggregated node at the current frame is connected to a previous node, all of its detection, i.e. bounding-box and camera ID, will inherit the same ID of that previous node. Otherwise, it will be assigned a new ID if there is no other previous node connected. Therefore, we can obtain predicted tracklet IDs corresponding to every detection at the current frame to accomplish inference.



(a) (b)

Figure 1: Tracklet ID Assignment. (a) In Graph Reconfiguration, we aggregate a list of bounding-boxes and camera IDs from its original connected component into a node (black solid node). (b) After Temporal Association, each node at time t and its associated detections will inherit ID from the previous node or be assigned to a new ID.

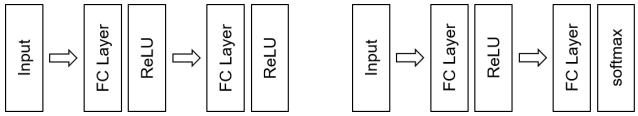
C. Network Details

Following [4], our ReST model contains five trainable MLPs (Table 1, Figure 2). $f_{FE}^v(\cdot)$ and $f_{FE}^e(\cdot)$ serve as initial feature encoders for nodes and edges to project the original features, e.g. appearance feature and geometry position, into a high-dimensional feature space. $f_{ME}^v(\cdot)$ and $f_{ME}^e(\cdot)$ are used in MPN. We encode the feature first, and then pass the message and update it. With the softmax layer appended at the end, $f_{CLS}(\cdot)$ outputs a confidence score between 0 and 1 from the enhanced edge feature.

Our graph model has about 154K parameters in to-

Network	Layer	Input	Output	Parameters
$f^v_{FE}(\cdot)$	FC + ReLU FC + ReLU	512 / 514 128	128 32	69K / 70K
$f^e_{FE}(\cdot)$	FC + ReLU FC + ReLU	4 / 6 8	8 6	94 / 110
$f^v_{ME}(\cdot)$	FC + ReLU FC + ReLU	38 64	64 32	4576
$f^e_{ME}(\cdot)$	FC + ReLU FC + ReLU	70 32	32 6	2470
$f_{CLS}(\cdot)$	FC + ReLU FC + softmax	6 4	4 1	33

Table 1: Details of each MLP network. The number before slash represents spatial graph, while the number behind represents temporal graph.



(a) Encoder (b) Classifier

Figure 2: MLP network architecture. (a) $f^v_{FE}(\cdot)$, $f^e_{FE}(\cdot)$, $f^v_{ME}(\cdot)$, $f^e_{ME}(\cdot)$ have the same structure. (b) f_{CLS} replaces ReLU with softmax to output a confidence score of each edge between 0 and 1.

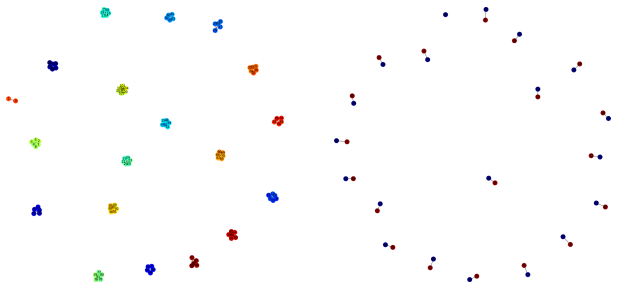
Setting	IDF1↑	MOTA↑
w/o splitting in both graphs	80.5	92.8
w/o splitting in spatial graph	84.3	92.4
w/o splitting in temporal graph	85.5	95.5
Ours (w/ full post-processing)	**91.6**	**97.0**

Table 2: Ablation of post-processing module on Wildtrack.

tal, which is a considerably light model compared with attention- or Transformer-based models [2, 5]. This makes ReST more suitable for real-world application scenarios.

D. Analysis on Post-processing Module

To validate the effectiveness of our post-processing module, we perform another ablation study on the post-processing module. In Algorithm 2, both spatial and temporal graphs perform pruning and splitting, while assigning tracklet ID will only perform in the temporal graph. Pruning and assigning ID are necessary and cannot be omitted, since pruning removes the edges and divides into several connected components representing different objects, and assigning ID is to output tracklet ID for evaluation. In practice, splitting is performed in both graphs to ensure that each connected component follows the specific constraints. As shown in Table 2, there is a significant decline in both metrics without the splitting in any graph or in both graphs.



(a) Spatial Graph (b) Temporal Graph

Figure 3: Graph visualization. Both graphs after their association stage are depicted. (a) In the spatial graph, each node represents one detection, colored by ground-truth label. (b) In the temporal graph, blue nodes represent aggregated nodes in temporal graph at previous frame, while red nodes represent aggregated nodes in spatial graph at current frame.

E. Graph Visualization

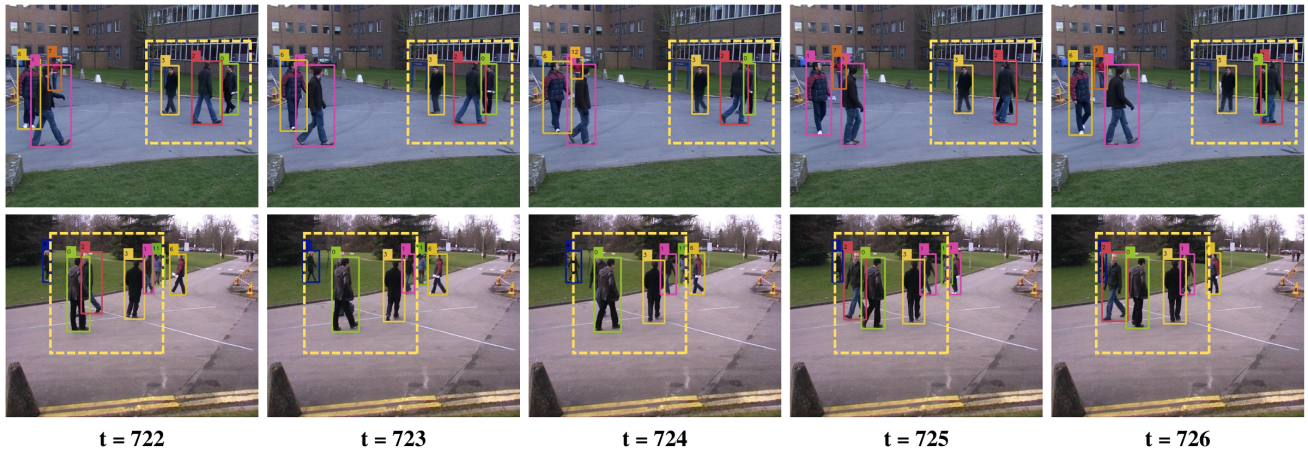
To better realize our graph model and prove the model robustness, we demonstrate the graph after the association stage in Figure 3. In Figure 3a, one node, colored by ground-truth label, represents one input detction at current frame. Our spatial graph perfectly associates every people across different views even in a crowded scene. In other words, we will not lose the information of occluded people, leading to fewer fragmented tracklets and ID switch errors. In Figure 3b, one node represents a list of detections that are aggregated in the Graph Reconfiguration stage. The connected nodes mean successful association between different frames, while the single nodes mean people who have just entered or left the scene. With the Graph Reconfiguration module, our view-invariant temporal graph becomes simple and focuses on associating nodes from different frames only.

F. Qualitative Results

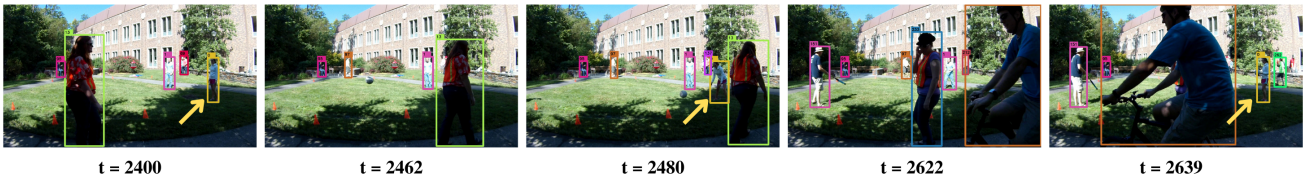
In this section, we show more cases fixing the fragmented tracklets problem due to occlusion in certain views. With the design of two-stages association, our ReST model leverages spatial consistency, recovering ID from different views. Specifically, an occluded person is usually visible in other views. We take advantage of this to fix the potential fragment and ID switch problems. As shown in Figure 4, no matter the short-term or long-term occlusion, we steadily track every people and do not lose their tracklet IDs.

G. Demonstration Video

We demonstrate our ReST tracker on Wildtrack [1] at `https://github.com/chengche6230/ReST`. To show spatial and temporal consistency, we simultaneously



(a) PETS09



(b) CAMPUS - Garden1

Figure 4: Qualitative results on PETS09 [3] and CAMPUS [6] to show our model's ability to recover fragmented tracklets. (a) ID:2 (red box) is occluded at time 723 to 724 in the second view, while he is still clearly visible in the first view. ID:0 (green box) is occluded at time 725 in the first view, but visible in the second view. Both cases are recovered and kept their IDs via our two-stages association. (b) Long-term consistency: ID:0 (yellow box) is occasionally occluded by people at time 2462 and 2622. Our model is able to maintain her ID in the long-term. As shown in the figure, we correct her ID at time 2480 and 2639.

show all frames from the 7 camera views and a bird's-eye view of their foot point. With the effective two-stages association, our predicted tracklet IDs are stable and consistent across views and frames.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, pages 5030–5039, 2018.

[2] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4870–4880, 2023.

[3] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance(IEEE)*, pages 1–6, 2009.

[4] Elena Luna, Juan C. SanMiguel, José M. Martínez, and Pablo Carballeira. Graph neural networks for cross-camera data association. *arXiv preprint arXiv:2201.06311*, 2022.

[5] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. DyGLIP: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13784–13793, 2021.

[6] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4256–4265, 2016.