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 $P_{c_{vi}}$ is based on camera calibration parameters and derived from perspective projection:

$$x_{img} = K[R t]x_{world} = Px_{world},$$  \hspace{1cm} (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 $3 \times 4$ projection matrix $P$ is reduced to be a $3 \times 3$ homography matrix $H$. Therefore, the position $p_{vi}$ on the ground plane can be calculated by

$$p_{vi} = H^{-1}x_{img},$$  \hspace{1cm} (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 $b_{vi}$ and camera ID $c_{vi}$. In Graph Reconfiguration stage, we save all bounding-boxes and their respective camera IDs within the same connected component of spatial graph $G^S_t$, 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.

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

<table>
<thead>
<tr>
<th>Network</th>
<th>Layer</th>
<th>Input</th>
<th>Output</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_{VE}(\cdot) )</td>
<td>FC + ReLU</td>
<td>512 / 514</td>
<td>128</td>
<td>69K / 70K</td>
</tr>
<tr>
<td>( f_{VE}(\cdot) )</td>
<td>FC + ReLU</td>
<td>4 / 6</td>
<td>8</td>
<td>94 / 110</td>
</tr>
<tr>
<td>( f_{ME}(\cdot) )</td>
<td>FC + ReLU</td>
<td>38</td>
<td>64</td>
<td>4576</td>
</tr>
<tr>
<td>( f_{ME}(\cdot) )</td>
<td>FC + ReLU</td>
<td>70</td>
<td>32</td>
<td>2470</td>
</tr>
<tr>
<td>( f_{CLS}(\cdot) )</td>
<td>FC + softmax</td>
<td>6</td>
<td>4</td>
<td>33</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Setting</th>
<th>IDF1↑</th>
<th>MOTA↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o splitting in both graphs</td>
<td>80.5</td>
<td>92.8</td>
</tr>
<tr>
<td>w/o splitting in spatial graph</td>
<td>84.3</td>
<td>92.4</td>
</tr>
<tr>
<td>w/o splitting in temporal graph</td>
<td>85.5</td>
<td>95.5</td>
</tr>
<tr>
<td>Ours (w/ full post-processing)</td>
<td><strong>91.6</strong></td>
<td><strong>97.0</strong></td>
</tr>
</tbody>
</table>

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

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.

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 detection. 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 person 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
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


