Video-based Person Re-identification with Spatial and Temporal Memory Networks Supplement

1. Hyperparameter

We use a grid search to set the sizes of the spatial and temporal memories, M and N, the pre-defined margin α in the memory spread loss, and the length of an input sequence, L. We randomly divide the training set of MARS [5] into two splits, and train and evaluate our model on each split (see Sec. 4.1 of the main paper for the details). We perform 5 trials and report the mean and standard deviation of rank-1 accuracy and mAP. The results are shown in Table 1, Table 2, and Table 3, respectively.

M	N	rank-1	mAP
5	5	94.1 ± 1.02	89.9 ± 1.17
	10	94.5 ± 0.79	89.6 ± 0.96
	20	94.4 ± 0.65	89.4 ± 1.29
10	5	**95.1** ± 0.66	**90.5** ± 0.46
	10	93.8 ± 1.15	90.0 ± 0.72
	20	94.4 ± 0.55	89.8 ± 0.96
20	5	94.7 ± 0.45	<u>90.3</u> ± 0.38
	10	<u>95.0</u> ± 0.61	**90.5** ± 0.57
	20	94.2 ± 0.97	89.8 ± 1.02

Table 1. Quantitative comparison for the sizes of the spatial and temporal memories. We measure rank-1 accuracy(%) and mAP(%). Numbers in bold indicate the best performance and underscored ones are the second best.

α	rank-1	mAP
0.1	94.0 ± 0.94	90.0 ± 0.66
0.3	**95.1** ± 0.55	**90.5** ± 0.46
0.5	<u>94.9</u> ± 0.96	<u>90.2</u> ± 0.64
0.7	94.0 ± 1.27	89.4 ± 1.15
1.0	93.9 ± 0.65	89.1 ± 0.98

Table 2. Quantitative comparison for the margin of the memory spread loss. We measure rank-1 accuracy(%) and mAP(%). Numbers in bold indicate the best performance and underscored ones are the second best.

Note that, in Table 1, as the sizes of the memories increase, the performance slightly degrades. During training, the spatial memory is shared by multiple IDs. Thus, the memory is discouraged to save information related to a certain ID (*i.e.*, useful priors). However, when the size of the spatial memory increases (*e.g.*, $M = 20$), we found that few memory items learn information related to a certain ID, and this may lead to performance drops considering that training/test sets of reID datasets do not share IDs. The memory spread loss enforces all items in the memory to be used, encoding dissimilar features with each other. This leads our memory to discover objects shared by multiple IDs other than background clutter and regard common person attributes, *e.g.*, black bags on MARS (see Fig. 1), as distracting scene details. Similarly, when the size of temporal memory increases, few memory items may store unnecessary temporal contexts, and this leads to slight performance drops on Rank-1/mAP.

L	rank-1	mAP
4	93.7 ± 1.15	<u>89.8</u> ± 0.25
6	**95.1** ± 0.66	**90.5** ± 0.46
8	<u>93.8</u> ± 0.57	89.6 ± 0.95
10	93.3 ± 0.57	87.8 ± 1.08

Table 3. Quantitative comparison for the length of an input sequence. We measure rank-1 accuracy(%) and mAP(%). Numbers in bold indicate the best performance and underscored ones are the second best.

Figure 1. Examples of person images that show one of the common person attributes, *i.e.*, a black bag, in MARS [5].

2. Implementation details

The encoder of STMN has three heads, which are used to extract person representations, query maps for the spatial memory, and query maps for the temporal memory, re-

spectively, from an input video frame. Each head has the same architecture as `conv5` layer of ResNet, except that the stride of the first residual block is set to 1. Note that, motivated by [1, 4], we replace the last ReLU layer [3] in the heads for extracting the query maps with a L2 normalization layer.

3. Comparison with the state of the art

We compare retrieval results for STE-NVAN [2] and STMN on MARS [5] in Fig. 2. We can see that STE-NVAN [2] is easily disturbed by spatial distractors (*e.g.*, playfield, concrete pavers), retrieving sequences of different persons from the query, but captured with the similar distractors and wearing analogous outfits as the query. On the contrary, STMN successfully retrieves person sequences with the same identity as the query, robust to the distractors.

4. Additional results

In this section, we provide more results for visual analysis on the spatial and temporal memories. We refine person representations from the encoder using the spatial memory. We verify the effectiveness of the refinement by comparing retrieval results for the original person representations $\mathbf{f}_i^{\mathrm{o}}$ and the refined ones $\mathbf{f}_i^{\mathrm{s}}$ in Fig. 3. We can see that, while the original representations retrieve wrong person sequences that are captured under similar background, the refined ones correctly retrieve sequences with the same identity as the query.

The refined representations are then fused using temporal attentions which are generated by the temporal memory. We visualize the temporal attentions with input sequences in Fig. 4. The examples show that the temporal memory generates the attentions which highlight discriminative frames of a given sequence, indicating that the memory allows fusing frame-level representations robust to temporal distractors.

References

[1] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *ICLR*, 2017. 2

[2] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *BMVC*, 2019. 2, 3

[3] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 2

[4] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *CVPR*, 2019. 2

[5] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*. 1, 2, 3, 4, 5

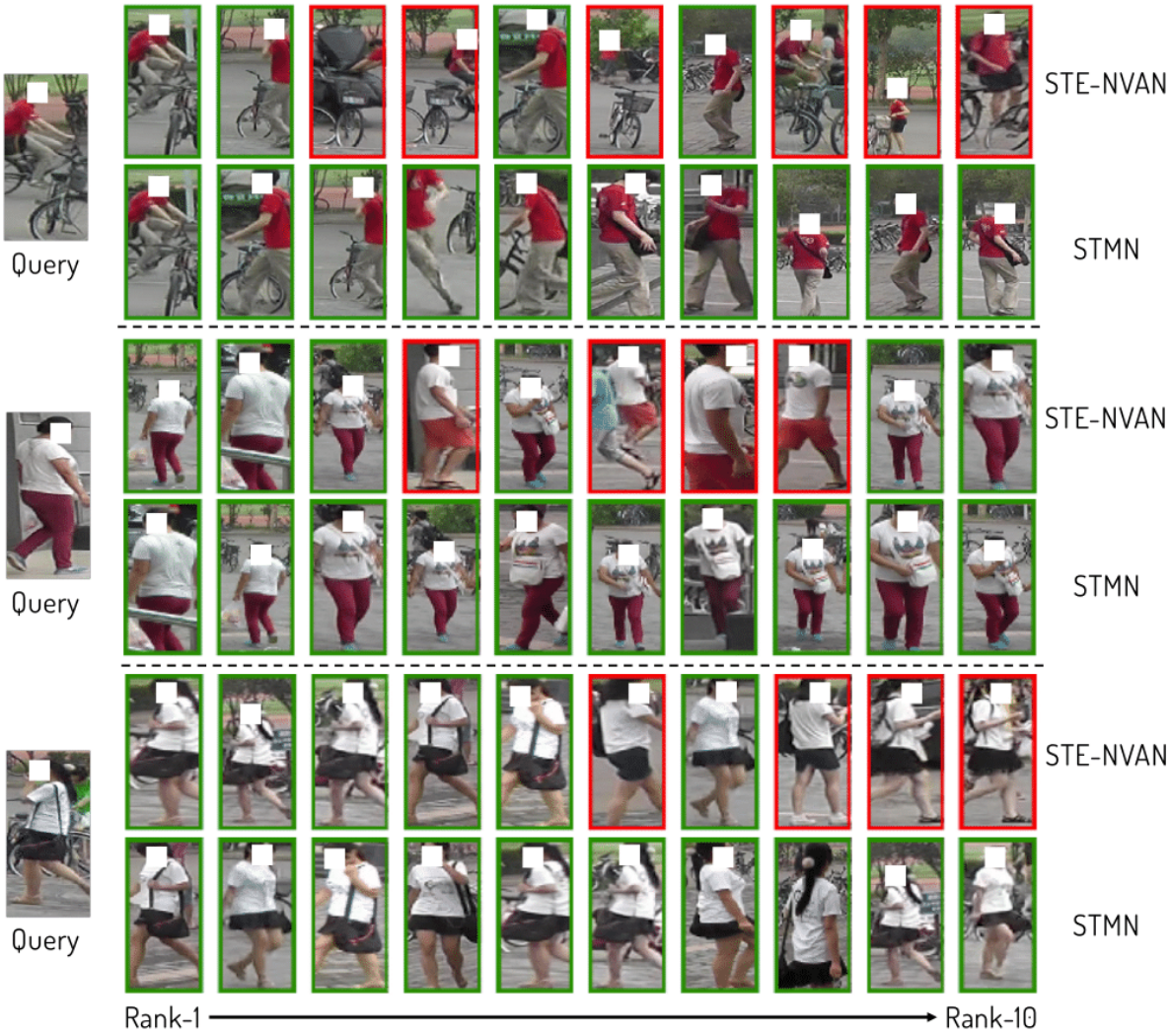

Figure 2. Comparison of top-10 retrieval results on the test split of MARS [5] between STE-NVAN [2] and STMN. Results with green boxes have the same identity as the query, while those with red boxes do not. We show the first frame of sequences for the purpose of visualization. (Best viewed in color.)

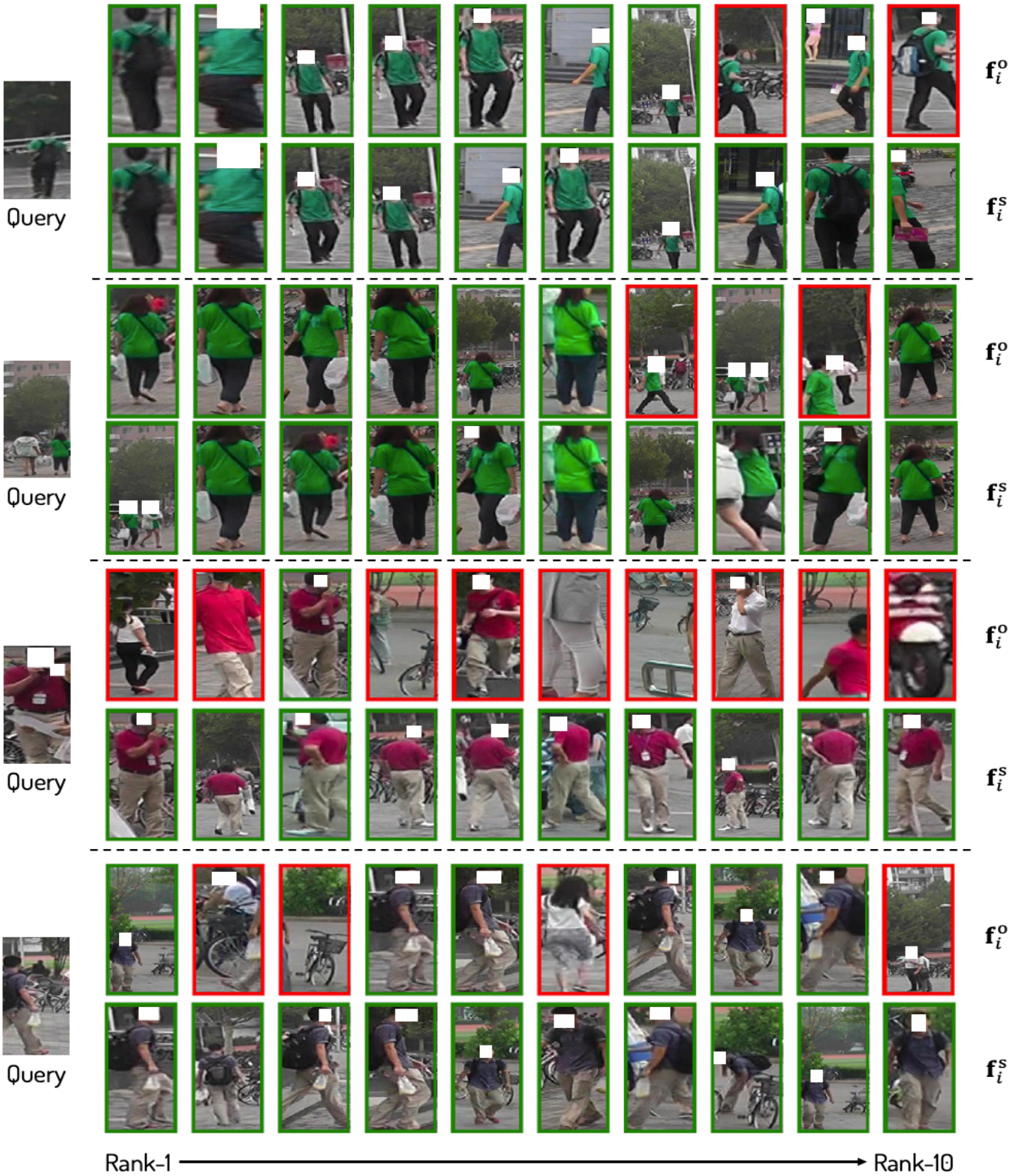

Figure 3. Comparison of top-10 retrieval results on the test split of MARS [5] using the original frame-level features $\mathbf{f}_i^o$ (top) and refined ones $\mathbf{f}_i^s$ (bottom). Results with green boxes have the same identity as the query, while those with red boxes do not. We show the first frame of sequences for the purpose of visualization. (Best viewed in color.)

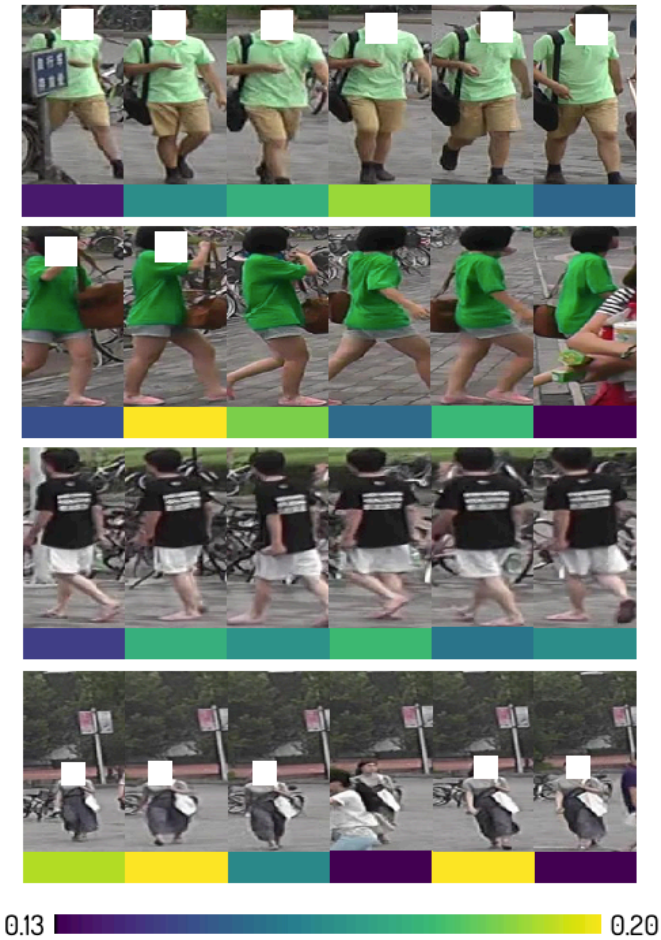

Figure 4. Examples of temporal attentions generated by the temporal memory on the test split of MARS [5]. (Best viewed in color.)