Supplementary Material for "Spatial-Temporal Graph Convolutional Network for Video-based Person Re-identification"

Jinrui Yang[1,3] , Wei-Shi Zheng[1,2,3*] , Qize Yang[1,3] , Yingcong Chen[4] , and Qi Tian[5]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4] The Chinese University of Hong Kong, China
[5]The Huawei Noah's Ark Lab, China
{yangjr27,yangqz}@mail2.sysu.edu.cn,wszheng@ieee.org,yingcong.ian.chen@gmail.com
,tian.qi1@huawei.com

1. Abstract

This supplementary material accompanies our main manuscript "Spatial-Temporal Graph Convolutional Network for Video-based Person Re-identification". Further analysis experiments on DukeMTMC-VideoReID [3, 2], and some visualization results are provided in this material.

2. More Analysis and Experiments

Due to space limitation in Section 4.5 and 4.6 in the main manuscript, we report the results of the analysis experiments on the key parameters of our method on MARS [4]. We provide the analysis experiments on DukeMTMC-VideoReID in this supplementary material as well.

2.1. The Number of GCN Layers in GCN Module

In our proposed model, the number of GCN layers in TGCN and SGCN are denoted as M and K, respectively. We carry out experiments to investigate the effect of the number of GCN layers by changing one of the GCN modules while freezing the other one.

The impact of the number of GCN layers in TGCN. In this experiment, we fix the number of GCN layers in SGCN (i.e., $K = 2$) then evaluate the performance of our model when $M = 3, 4, 5, 6, 7$. From Figure 1 (a), we can see that the best Rank-1 and the best mAP are 97.44% and 95.94% respectively when $M = 4$. The result outperforms the state-of-the-art methods and the baseline model by a large margin.

The impact of the number of GCN layers in SGCN. Similarly, we fix the number of GCN layers in SGCN (i.e., $M = 4$) then evaluate the performance of our model when $K = 1, 2, 3, 4$. As shown in Figure 1 (b), when $K = 3$, the model achieves the best performance. The Rank-1 is 97.44% and the mAP is 96.00%.

As shown in Figure 1, the performance of STGCN is always higher than the baseline model (i.e., 94.08%/96.01% in mAP/Rank-1) and the state-of-the-art methods, although the number of GCN layers can affect the performance of the model.

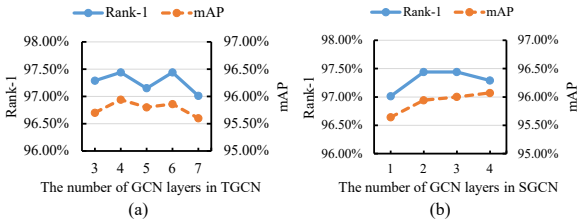


Figure 1. (a) Analysis on the number of GCN layers in TGCN. (b) Analysis on the number of GCN layers in SGCN. We carry out these experiments on the DukeMTMC-VideoReID dataset.

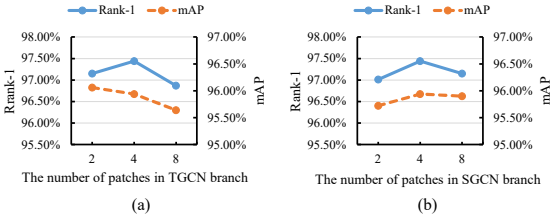


Figure 2. (a) Analysis on the number of patches in TGCN. (b) Analysis on the number of patches in SGCN. We carry out these experiments on the DukeMTMC-VideoReID dataset.

2.2. Analysis on the Number of Patches in GCN Module

The number of nodes in the graph (i.e., the number of patches) is another key parameter of GCN. For convenience,

*Corresponding author

we denote the number of patches of each frame in TGCN and SGCN are P^t and P^s, respectively.

The impact of the number of patches in TGCN. In this experiment, we fix $P^s = 4$ and evaluate the results when $P^t = 2, 4, 8$. From the Figure 2 (a), we can see that the model has the best performance when p^t=4. The Rank-1 is 97.44% and the mAP is 95.94%.

The impact of the number of patches in SGCN. Similarly, we fix $P^t = 4$ and evaluate the results when $P^s = 2, 4, 8$. As shown in Figure 2 (b), the model has the best performance when p^s=4. The Rank-1 is 97.44% and the mAP is 95.94%.

As shown in Figure 2, the performance of STGCN is always higher than the baseline model (i.e., 94.08%/96.01% in mAP/Rank-1) and the state-of-the-art methods, although the number of patches can affect the performance of the model.

2.3. Summary

Combined with Section 4.6, it can be seen that the role of parameters of our proposed method on DukeMTMC-VideoReID dataset is basically the same as on MARS dataset. It always has obvious improvement under different parameters and different datasets, which shows our proposed method is robust and effective.

3. Visualization

Visualization of class activation maps We visualise the class activation maps (CAMs) in Figure 3 by using Grad-CAM [1].

From Figure 3 (a) and (b), it can be seen that the baseline model only pays attention to few local body regions. However, it is obvious that our proposed method can focus on more discriminative pedestrian body parts, which can be used to learn the structural information of pedestrian by modeling the spatial relations of pedestrian patches.

Meanwhile, as Figure 3 (c) and (d) are illustrated, we can observe that the occlusion has a serious impact on the baseline model. Specifically, for Figure 3 (d), it is clear that the baseline model almost can not pay attention to any useful parts because the disturbance of occlusion. But our proposed model still can focus on these unoccluded parts which are discriminative regions. In addition, for Figure 3 (c), we can see that there are both unoccluded frames and occluded frames in this image sequence. For these unoccluded frames, our proposed method is slightly better than the baseline model. But for these occluded frames, it is clear that the baseline model can not focus on useful body parts. On the contrary, our proposed method can effectively and accurately focus on unoccluded pedestrian body parts, which can alleviate occlusion problem.

Retrieval results analysis As shown in Figure 4, we can see that the top 5 results of our proposed method is all matching. However, the Rank-5 result of the baseline model is a wrong match due to the problem of occlusion.

From the Figure 5, we can see the appearance of the query image sequences and the gallery image sequences are very similar. It is also clear that the top 5 results of our proposed method is all matching. But the Rank-1 and Rank-5 results of the baseline model are disturbed by the samples of other identities with similar appearance.

Summary From the class activation maps and retrieval results above, we can know that they both can prove our proposed method indeed alleviate the problem of similar appearances of different identities and occlusion problem by modeling the spatial and temporal relations of patches.

References

[1] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[2] Xiaogang Wang and Rui Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. Springer, 2014. 1

[3] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018. 1

[4] 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 *European Conference on Computer Vision*, pages 868–884. Springer, 2016. 1

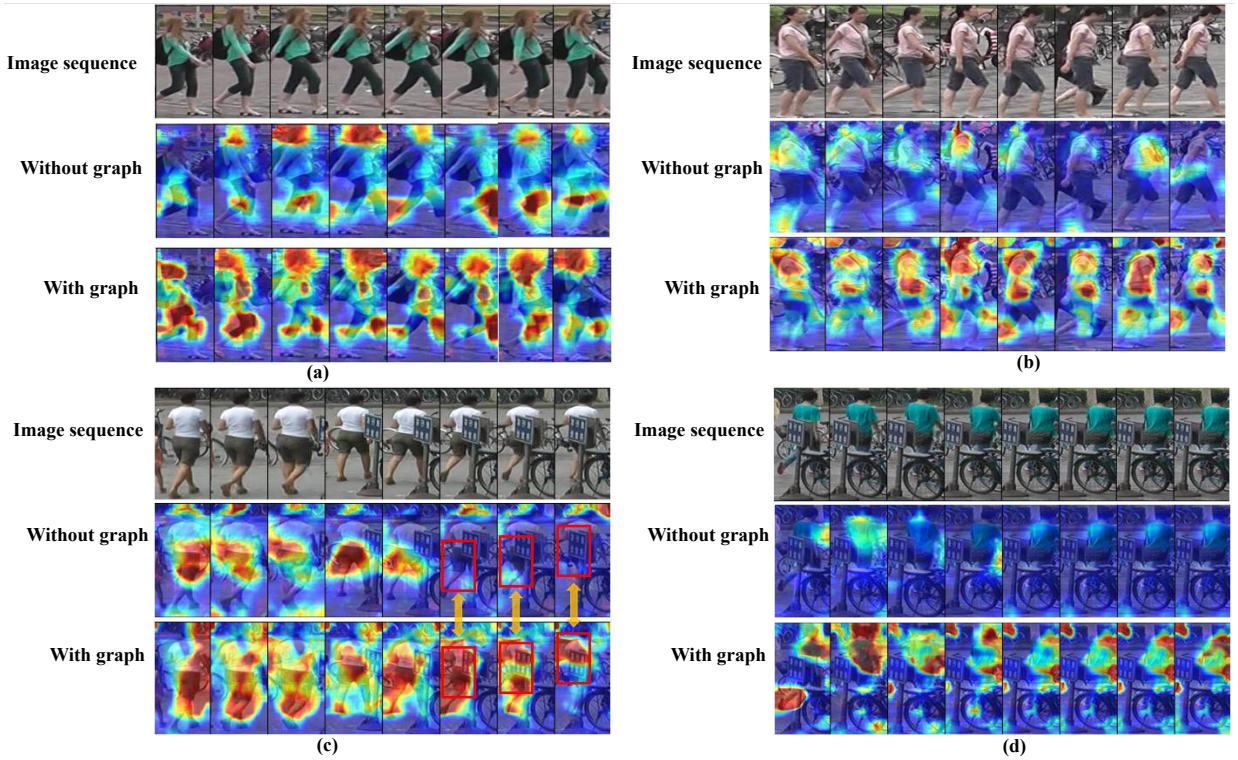
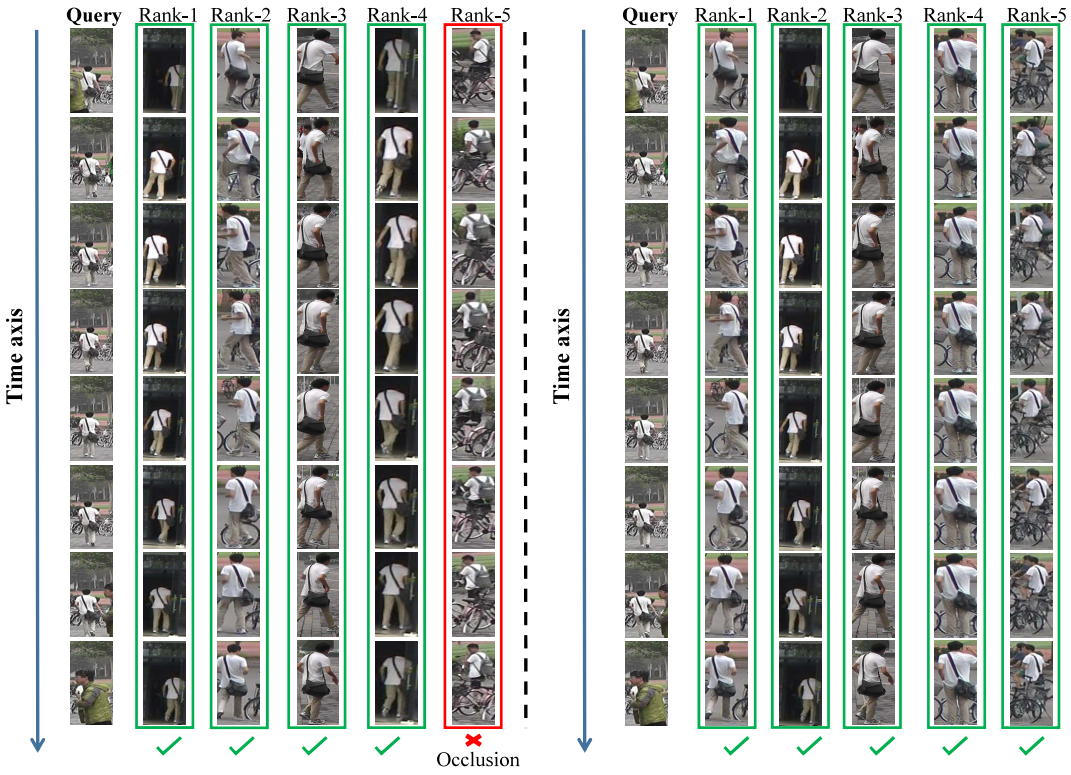


Figure 3. The visualization of the class activation maps (CAMs). Here we show the visualization of the four different image sequences. The visualization of these class activation maps are denoted as (a), (b), (c) and (d) respectively. For each subfigure (a), (b), (c) and (d), the first row is the original image sequences from MARS. The second row is the class activation maps of the baseline model. The third row is the class activation maps of our proposed model. Compared to the baseline model, our proposed method can fully make use of the relations of patches to help distinguish different pedestrians with similar appearance and become more robust to occlusion. **Best viewed in color.**



(a) Baseline model

(b) Our proposed method

Figure 4. (a) and (b) are the top 5 retrieval results of the baseline model and our proposed method in the MARS dataset, respectively. Different from image-based Re-ID, the query and gallery both are image sequences in video-based Re-ID. The green box is the correct matched result and the red box is the wrong matched result. Each column represents an image sequence. We can see that the baseline model returns a wrong Rank-5 match due to the problem of occlusion. However our proposed method still returns the correct match in the presence of occlusion, which shows our proposed method is robust to occlusion by modeling the temporal relations of whole patches in videos. **Best viewed in color.**

Figure 5. (a) and (b) are the top 5 retrieval results of the baseline model and our proposed method in the MARS dataset, respectively. Different from image-based Re-ID, the query and gallery both are image sequences in video-based Re-ID. The green box is the correct matched result and the red box is the wrong matched result. Each column represents an image sequence. We can see that the baseline model does not distinguish well between different pedestrians with similar appearance. Thus the baseline model returns the wrong Rank-1 and Rank-5 results. But our proposed method is all matching, which can illustrate our proposed method can effectively distinguish different pedestrians with similar appearance by learning the spatial relations of pedestrian patches. **Best viewed in color.**