Supplementary Material: Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification

Zhizheng Zhang[1] Cuiling Lan[2] Wenjun Zeng[2] Zhibo Chen[1]
[1]University of Science and Technology of China [2]Microsoft Research Asia
zhizheng@mail.ustc.edu.cn {culan,wezeng}@microsoft.com chenzhibo@ustc.edu.cn

1. More Experimental Details

1.1. Details of the Datasets

MARS [9] is a large-scale benchmark dataset for video-based person reID. It is captured by six cameras deployed in a university and contains 17503 tracklets corresponding to 1261 identities in total, of which 625 identities are used for training while the other 636 identities are used for testing. Besides, the additional 3248 tracklets are taken as distractors. The average number of frames per tracklet is 59. The bounding boxes for the target person are detected by Deformable Part Model (DPM, [2]) and tracked by the GMCP tracker [8].

iLIDS-VID [7] contains 600 video tracklets of 300 identities captured by 2 cameras. The length of sequences ranges from 23 to 192 frames with an average number of 73 frames. The bounding boxes are manually annotated.

PRID2011 [4] consists of 400 video tracklets of 200 identities from 2 cameras. The sequence length varies from 5 to 675. As in the previous works [9, 10, 5], we select video sequences with more than 21 frames for training and testing, leading to 178 identities. The bounding boxes are manually annotated.

1.2. Details of the Evaluation Metrics

The Cumulative Matching Characteristic (CMC) and the mean average precision (mAP) are used for evaluation. We use CMC to represent the number of true matching samples within the first k ranking results, which indicates the accuracy of person retrieval. Following the common practices, we report Rank-1, Rank-5, Rank-10, and Rank-20 accuracy, and also use mAP to evaluate the model performance in MARS. We obtain the average precision (AP) for each query and calculate the mean value of AP across all queries to obtain the mAP accuracy.

1.3. Implementation Details

For all the reported models, the backbone network is pre-trained on ImageNet [1] and both identification (cross entropy) loss with label smoothing [6] and triplet loss with hard mining [3] are used. We use Adam optimizer with a weight decay of 5×10^{-4}. We warm up the models for 20 epochs with a linear growth learning rate from 8×10^{-6} to 8×10^{-4}. Then, the learning rate is decayed by a factor of 0.5 for every 40 epochs. We observe that the models converge after a training of 320 epochs and we use them for testing. All our models are implemented with PyTorch and trained on two P40 GPUs.

Table 1: Performance (%) of using single granularity reference-aided attentive feature aggregation (SG-RAFA) at different granularities. "SG" denotes "Single-Granularity" and "MG" denotes "Multi-Granularity". N denotes the number of granularities. *G-1st* denotes the finest granularity with the spatial resolution of 16×8 for F_{all} and F_R, while *G-4th* denoting the coarsest granularity with the spatial resolution of 2×1 for F_{all} and F_R.

Models	MARS				
	mAP	R-1	R-5	R-10	R-20
Baseline	82.1	85.9	95.1	96.5	97.3
SG-RAFA (*G-1st*)	84.9	88.4	96.6	97.6	98.5
SG-RAFA (*G-2nd*)	84.6	87.5	96.5	97.2	98.0
SG-RAFA (*G-3rd*)	84.2	87.4	95.8	97.1	97.9
SG-RAFA (*G-4th*)	83.2	86.7	95.3	96.7	97.7
MG-RAFA (*N=4*)	**85.9**	**88.8**	**97.0**	**97.7**	**98.5**

2. More Ablation Studies

2.1. Performance of Different Granularities

In the paper, we have demonstrated the effectiveness of using multiple granularities and compared the corresponding performance with that using a single granularity in Section 4.3.1. For the single granularity setting in our paper, we use the finest granularity, *i.e.*, *G-1st*. Here, we study the performance of using other granularity alone, *e.g.*, *G-2nd*, *G-3rd*, *G-4th*, respectively, and show the results in Table 1. Here, *G-1st* denotes the finest granularity with the spatial resolution of 16×8 for F_{all} and F_R, while *G-4th* denoting the coarsest granularity with the spatial resolution of 2×1

Table 2: The ablation study of different pooling strategies to obtain coarse granularity features. "Max-Pooling" denotes max pooling strategy while "Avg-Pooling" denoting average pooling strategy.

Models	MARS				
	mAP	R-1	R-5	R-10	R-20
Baseline	82.1	85.9	95.1	96.5	97.3
Max-Pooling	84.1	87.3	95.9	97.4	98.0
Avg-Pooling (Ours)	**85.9**	**88.8**	**97.0**	**97.7**	**98.5**

for F_{all} and F_R. Note that S (which denotes the number of splits(groups) along the channel dimension for masking attention on each split respectively) is set as four for these schemes.

From Table 1, we observe that: (1) our proposed *MG-RAFA (N=4)* is superior to all the single-granularity schemes, which indicates that the exploration of different level semantics at different granularities are complementary and the joint exploration is effective. (2) All the *SG-RAFA* schemes of different granularities outperform the scheme *Baseline* consistently, which demonstrates the effectiveness of our proposed reference-aided attentive feature aggregation solution.

2.2. Study on Pooling Strategies

To obtain the features at coarse granularity, we perform average pooling on spatial feature positions (as described in Section 3.3 in our paper). We study the influence of different pooling strategies, including max pooling operation (denoted by "Max-Pooling"), and average pooling operation (denoted by "Avg-Pooling"). The results are shown in Table 2. We observe that average pooling strategy outperforms max pooling by 1.8% in mAP and 1.5% in Rank-1 respectively. Intuitively, average pooling can keep more information than max pooling and is robust to noise.

2.3. Multi-Granularity Design for F_{all} and F_R

In our proposed scheme *MG-RAFA*, for each coarse granularity, both the spatial temporal feature nodes F_{all} and reference nodes F_R are spatially average pooled to have a lower resolution. This ensures that the relation modeling and attention learning are at the same semantics level for each granularity. We validate the effectiveness of modeling at the same semantics levels in Table 3. Here, $SG\text{-}F_{all}$ denotes the scheme that we only generate multi-granularity features for the S-RFNs (*i.e.* F_R) while using the original granularity features for the spatial temporal features F_{all}. We compare $SG\text{-}F_{all}$ to our final scheme *MG-RAFA* which generates multi-granularity features for F_{all} (denoted by $MG\text{-}F_{all}$). We observe that $MG\text{-}F_{all}$ outperforms $SG\text{-}F_{all}$ by 2.6% in mAP and 2.4% in Rank-1, demonstrating the effectiveness of our matched multi-granularity design.

Table 3: The comparison of using single-granularity frame-level features (denoted by "$SG\text{-}F_{all}$") and using multi-granularity frame-level features (denoted by "$MG\text{-}F_{all}$") in the feature aggregation with multi-granularity S-RFNs.

Models		MARS				
		mAP	R-1	R-5	R-10	R-20
Baseline		82.1	85.9	95.1	96.5	97.3
SG-F_{all}		83.3	86.4	95.1	96.4	97.2
MG-F_{all}	(MG-RAFA)	**85.9**	**88.8**	**97.0**	**97.7**	**98.5**

3. More Visualization

In this section, we visualize the attention masks learned by our proposed multi-granularity reference-aided attentive feature aggregation (MG-RAFA) scheme across different frames and granularities on four video sequences in Table 1. Note that in our paper, we only partially show the visualization results due to space limitation.

From Figure 1, we have several observations. (1) The learned attention tends to focus on different semantic regions from different frames, which gets rid of a lot of repetitions (redundancy). (2) The learned attention is able to select the better represented areas (*e.g.*, at the $2nd$ granularity (*G-2nd*) of the right-bottom example, the shoe region in the $8th$ frame is clearly visible and thus has high response) and exclude the interferences (*e.g.*, at all granularities of the left-top example, the left regions are not attenteded because there are occlusions). (3) It seems our model captures different semantics at different granularities, which tends to capture more details at finer granularities and larger body parts at coarser granularities. At the coarsest granularity, *i.e. G-4th*, it seems the attention plays a role of selecting some frames to exclude redundancy. We believe our reference-aided attention modeling is an effective method to capture and learn discriminative spatial and temporal representation.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Pedro F Felzenszwalb, David A McAllester, Deva Ramanan, et al. A discriminatively trained, multiscale, deformable part model. In *CVPR*, volume 2, page 7, 2008. 1

[3] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1

[4] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 1

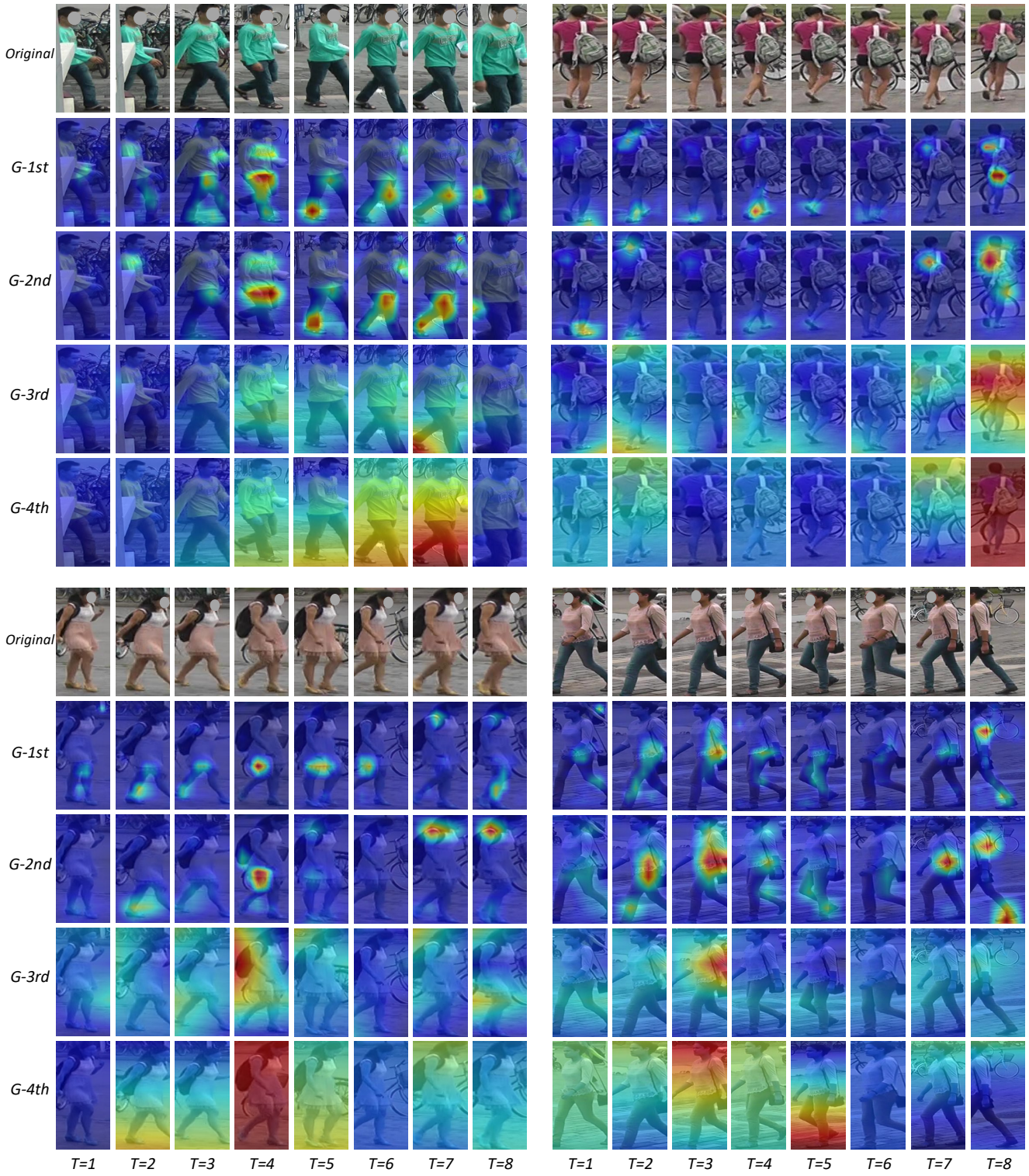

Figure 1: Visualization for the learned attention masks of our proposed MG-RAFA across different frames (columns) and granularities (rows). *G-1st* to *G-4th* denote the *1st* to the *4th* granularities, and the corresponding spatial resolutions of the attention masks for each frame are 16×8, 8×4, 4×2, 2×1, respectively. For different granularities, we rescale the attention masks of different spatial resolutions to the same resolution for visualization.

[5] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 5790–5799, 2017. 1

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 1

[7] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014. 1

[8] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *ECCV*, pages 343–356. Springer, 2012. 1

[9] 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*, pages 868–884. Springer, 2016. 1

[10] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person reidentification. In *CVPR*, pages 4747–4756, 2017. 1