Supplementary Material: Relation-Aware Global Attention for Person Re-identification

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

1. More Implementation Details

Networks. We show our network architecture (with ResNet-50 as backbone network and four RGA-SC attention modules added after the four convolutional blocks) in Table 1. Given an input image, we obtain a feature vector $\mathbf{z}$ of 2048-dimension after the global average pooling, then input this feature vector $\mathbf{z}$ to a Batch Normlization layer and get the feature vector $\hat{\mathbf{z}}$. Following the common practice, both triplet loss with hard mining [3] and classification loss (cross entropy loss with label smoothing [6, 9]) are adopted in training, which are added on $\mathbf{z}$ and $\hat{\mathbf{z}}$ respectively with weights of 1.0 and 1.0. The classifier consists of a Fully Connected layer (FC) with Softmax activation, and the dimension of its output is equal to the number of identities in the training set. For the triplet loss with hard mining [3], we sample P=16 identities and K=4 images for each identity as a mini-batch with the margin parameter set to 0.3. We in general use the l_2 normalized feature vector of $\mathbf{z}$ for feature l_2 distance calculation with in testing (on CUHK03, MSMT17). For Market-1501, similar to [5], we use the l_2 normalized feature vector of $\hat{\mathbf{z}}$ for feature distance calculation in testing, which gives performance a little better than that using $\mathbf{z}$ (the difference is small, *i.e.*, 0.7% in mAP, 0.2% in Rank-1 for our final scheme).

Training. For all the re-id models, we initialize the residual blocks with ImageNet [1] pre-trained weights and train the entire model for 600 epochs in total by adopting the Adam optimizer. During the training, we first warm up 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 before the 360th epoch. All our models are implemented on PyTorch and trained on a single P100 GPU.

Table 1: Our detailed architecture (built based on ResNet-50) with attention modules for the re-id task. The output sizes are denoted as: height × weight × number of channels, where the spatial resolution of the input (image) is 256×128. The convolution parameters are denoted by kernel height × kernel width, number of filters, convolution stride, respectively. N_{id} denotes the number of identities for the classification loss. Note that we add our attention modules RGA-SC or other attention modules in this way for comparisons in our ablation studies in the paper.

Layer Name	Output Size	Network Architecture
conv1	$128 \times 64 \times 64$	7×7, 64, stride 2
bn1	$128 \times 64 \times 64$	Batch Normalization + ReLU
conv2_x	$64 \times 32 \times 64$	3×3, max pool, stride 2
	$64 \times 32 \times 256$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
attention1	$64 \times 32 \times 256$	Attention Module $\times$ 1
conv3_x	$32 \times 16 \times 512$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
attention2	$32 \times 16 \times 512$	Attention Module $\times$ 1
conv4_x	$16 \times 8 \times 1024$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
attention3	$16 \times 8 \times 1024$	Attention Module $\times$ 1
conv5_x	16×8	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
attention4	$16 \times 8 \times 2048$	Attention Module $\times$ 1
pooling	$1 \times 1 \times 2048$	Global Average Pooling
bn	$1 \times 1 \times 2048$	Batch Normalization
fc (classifier)	N_{id}	Fully Connected Layer

2. Detailed Introduction for Datasets

CUHK03 [4] consists of 1,467 pedestrians from six non-overlapped cameras. This dataset provides both manually labeled bounding boxes and DPM-detected bounding boxes from 14,097 images. The new training/testing protocol of [12, 11, 2] is used where 767 identities are used for training.

Market1501 [10] consists of 1,501 identities with 751 for training and the rest for testing. It has 12,936 training images, 3,368 query images and 19,732 gallery images.

MSMT17 [7] is a newly released large dataset which

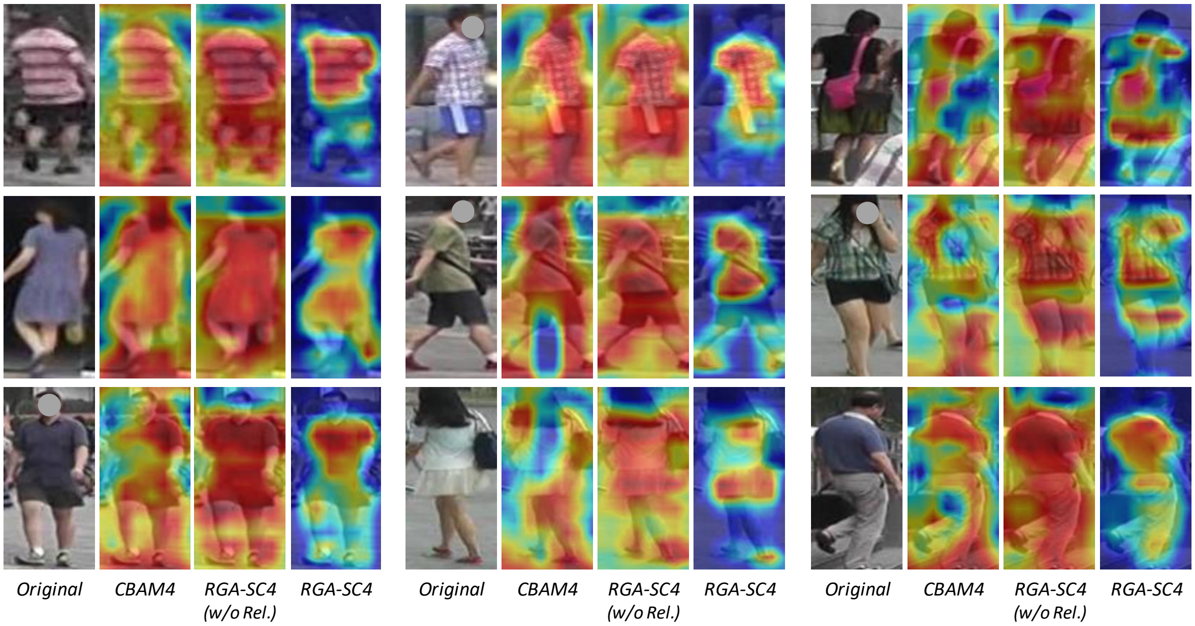

Figure 1: Visualization of the spatial attention maps learned by different attention modules. The four columns of each sub-image correspond to the original images, CBAM modules, RGA-SC without the relations adopted (referred to as *RGA-SC4 w/o Rel.* in short), and RGA-SC modules, respectively. Note that we visualize the learned attention maps of the fourth attention module (in the position of "attention4" as illustrated in Table. 1) for all the schemes.

consists of 4101 identities from 15 cameras (including 12 outdoor cameras and 3 indoor cameras). It provides a total of 126441 images with different weather conditions and with bounding boxes annotated.

3. Where to Plug-in RGA-SC Module?

We compare the cases of adding a single RGA-SC module to different residual blocks, and to all the four blocks (*i.e.*, residual blocks of conv2_x, conv3_x, conv4_x, conv5_x) of our ResNet-50 baseline. Table 2 shows the detailed results. For each residual block, the module is added following its last layer. The improvement (in mAP) of RGA-SC is significant on conv3_x, conv4_x, is smaller on conv2_x, and is the smallest on conv5_x. For the lower layers conv2_x, only limited low-level semantics can be captured. For the higher layers conv5_x, there is less room for feature improvement. Thus, adding RGA-SC to the conv3_x or conv4_x outperforms that to conv2_x or conv5_x. When RGA-SC is added to all four blocks, our scheme achieves the best performance and outperforms the baseline by **6.6%** and **7.5%** in Rank-1 and mAP respectively on CUHK03.

4. More Visualization Results

In this section, we compare the attention maps of the fourth attention module (in the position of "attention4" as illustrated in Table 1) for different attention approaches in Figure 1, and visualize the learned spatial attention maps corresponding to the four RGA-SC modules (which are added to different residual blocks) in Figure 2.

Table 2: Performance (%) comparisons of adding RGA-SC modules to different residual blocks of ResNet50 baseline.

Model	CUHK03 (L)		Market1501	
	R1	mAP	R1	mAP
Baseline	73.8	69.0	94.2	83.7
ResBlock conv2_x	76.5	71.9	95.5	86.3
ResBlock conv3_x	78.1	73.4	95.6	86.5
ResBlock conv4_x	78.6	74.7	95.8	87.2
ResBlock conv5_x	74.0	71.2	94.7	85.5
ResBlock conv2,3,4,5_x	**80.4**	**76.5**	**95.8**	**88.1**

Attention Learned by Different Approaches. In order to better understand the effectiveness of our proposed relation-aware global attention, we show the comparisons of the learned spatial attention maps of CBAM [8], RGA-SC without relation features adopted (*RGA-SC w/o Rel.*), and our final scheme (*RGA-SC*). As shown in Figure. 1, we find that the attention results from *CBAM* and *RGA-SC w/o Rel.* can also attend to human bodies but are not concentrated well on the bodies, *i.e.*, with low spatial precision. Benefitting from the global relation exploration, the attention learned from our final model *RGA-SC* can more precisely focus on the discriminative regions.

Spatial Attention of Different Positions. To better understand our spatial attention, we visualize the spatial atten-

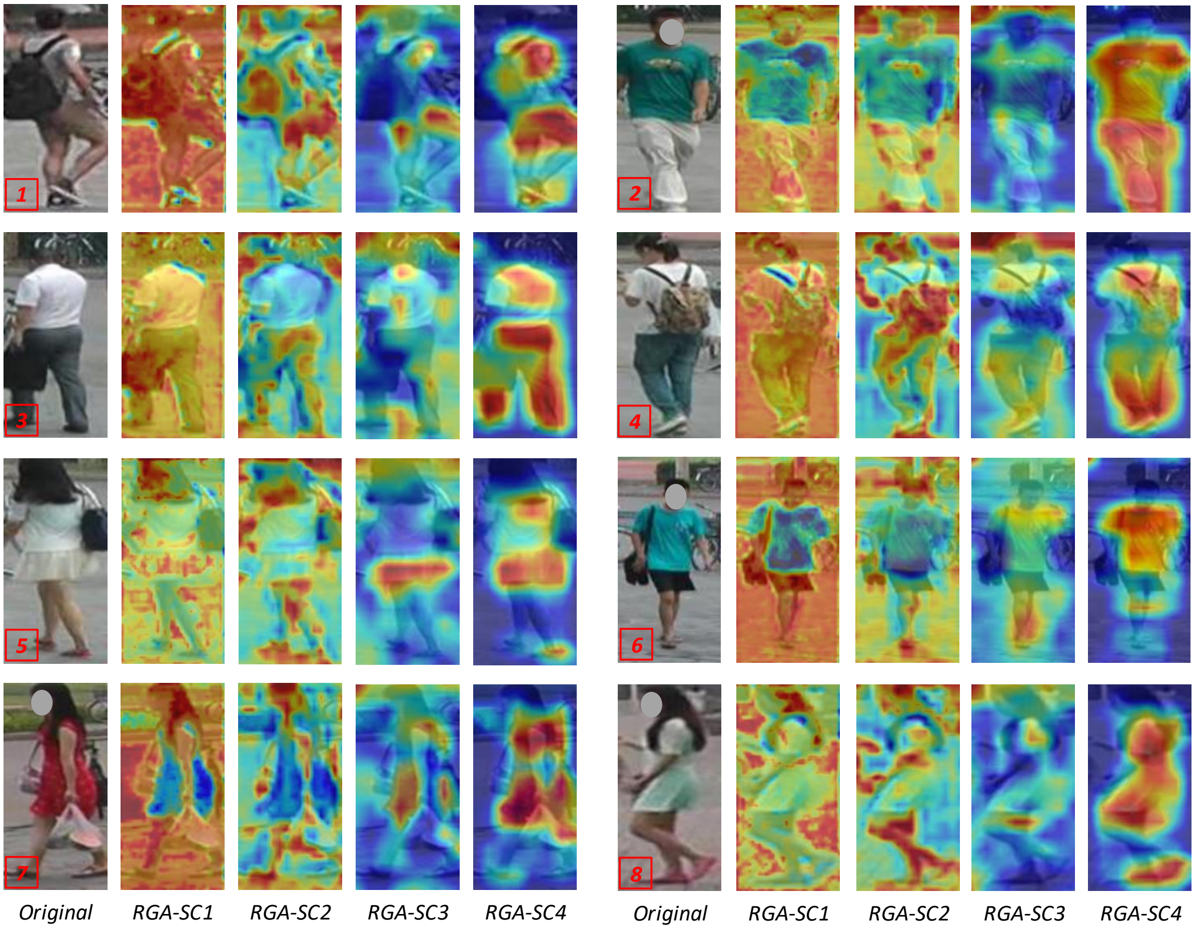

Figure 2: Visualization of the spatial attention maps generated by the four RGA-SC modules (*RGA-SC1* to *RGA-SC4*) which are added after different residual blocks, respectively. They correspond to attention 1, 2, 3, 4 in Table 1, respectively.

tion maps generated by the four RGA-SC attention modules corresponding to the four residual blocks (see Table 1) respectively in Figure. 2 (as marked by *RGA-SC1*, *RGA-SC2*, *RGA-SC3*, and *RGA-SC4*, respectively). As shown in Figure 2, the spatial attention maps present clustering-like patterns. From *RGA-SC1* to *RGA-SC4*, it seems that the semantics are captured progressively as the depth increases. For the attention map of *RGA-SC4* (which is followed by the global average pooling to obtain the re-id feature), the attention map has large responses on the discriminative human body regions, which successfully captures the high level semantics. For the lower layer like *RGA-SC2*, the attention map presents clustering-like patterns at finer granularity levels. The parts/regions with similar semantics are likely to have similar attention responses. For example on *RGA-SC2*, the leg region has similar response intensity for the 3^{rd} person image example and the backpack region has similar response intensity on the 4^{th} person image example. Moreover, for the lower layers, the purpose of attention is to enlarge the difference of different features rather than directly emphasize the important features with a large weight. In other word, it could mask the important feature by a small weight (small attention value) while masking the unimportant feature by a large weight to differentiate them.

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] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399*, 2018. 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] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1

[5] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, pages 0–0, 2019. 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] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 1

[8] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 2

[9] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 1

[10] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1

[11] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *TCSVT*, 2018. 1

[12] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Reranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 1