Mask-guided Contrastive Attention Model for Person Re-Identification Supplementary Materials

Chunfeng Song[1,3] Yan Huang[1,3] Wanli Ouyang[4] Liang Wang[1,2,3]
[1]CRIPAC & NLPR [2]CEBSIT & CASIA
[3]University of Chinese Academy of Sciences (UCAS)
[4]University of Sydney
{chunfeng.song, yhuang, wangliang}@nlpr.ia.ac.cn wanli.ouyang@sydney.edu.au

The supplementary document provides more details of proposed Mask-guided Contrastive Attention Model (MGCAM). Firstly, we provide the detailed network structure of proposed MGCAM, as shown in Table 1. Secondly, we show the CMC curves comparing with the state-of-the-art methods [5, 2] and comparing with different distance metrics on MARS [5]. The results of compared methods are shown in Figure 1. In Figure 2, the results of different distance metrics with our MGCAM-Siamese are reported. Thirdly, we show the Rank-1 and mAP accuracy maps between camera pairs in Figure 3 and Figure 4 on MARS. Finally, we also provide some segmentation examples from the used three datasets in Figure 5. **All of the masks used in our experiments will be released upon request.**

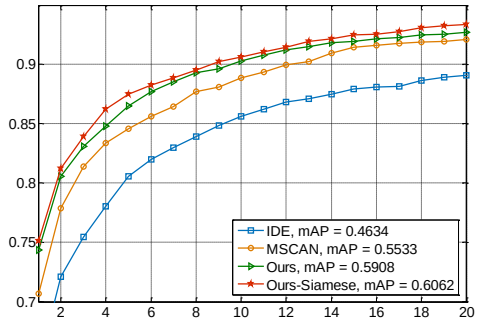

Figure 1. CMC curves of the compared methods on the MARS [5] dataset. We compare proposed methods with two state-of-the-art methods, including IDE [5] and MSCAN-body [2] with RGB-M as its inputs. It is obvious that our methods outperform the compared methods with a clear margin. All the methods are using the XQDA distance metric.

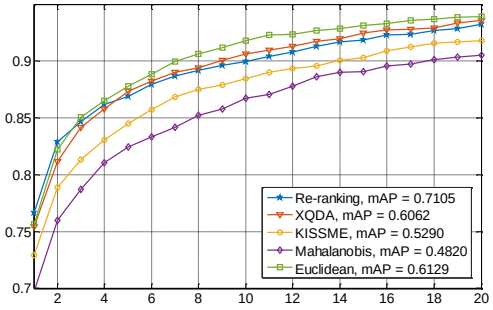

Figure 2. CMC curves of different distance metrics on the MARS [5] dataset. Experiments are conducted with four classic distance metrics, including the Euclidean distance, Mahalanobis distance, XQDA [4], KISSME [1] and the recently proposed Re-ranking methods [7]. Though the Re-ranking metric performs best in rank-1 accuracy, it becomes worse in rank-3 to rank-20. All results are evaluated with our MGCAM-Siamese method.

References

[1] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1

[2] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 1, 2

[3] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 3

[4] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 1

[5] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 1, 2, 3

[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 3

[7] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 1

stage	layer	dilation	kernel	pad	#filters	output	stream
-	input	-	-	-	-	4x160x64	1
	conv0	1	5x5	2	32	32x160x64	
	pool0	-	2x2	-	-	32x80x32	
1	conv1	1/2/3	3x3	1/2/3	32/32/32	96x80x32	
	pool1	-	2x2	-	-	96x40x16	
2	conv2	1/2/3	3x3	1/2/3	32/32/32	96x40x16	
	att-body	1	3x3	1	1	1x40x16	-
	att-bkgd	-	-	-	-	1x40x16	
	pool2	-	2x2	-	-	96x20x8	3
3	conv3	1/2/3	3x3	1/2/3	32/32/32	96x20x8	3
	pool3	-	2x2	-	-	96x10x4	
4	conv4	1/2/3	3x3	1/2/3	32/32/32	96x10x4	
	pool4	-	2x2	-	-	96x5x2	
-	fc1	-	-	-	-	128	
	fc2	-	-	-	-	#ID	

Table 1. The detailed model architecture of proposed MGCAM. Different from the body-version of MSCAN [2], MGCAM has three streams after the 2nd stage, i.e., the full-image stream, body stream and background stream. The three streams are with the same structures, except that the body and background streams are from the body-aware and background-aware attention features after the 2nd stage, respectively.

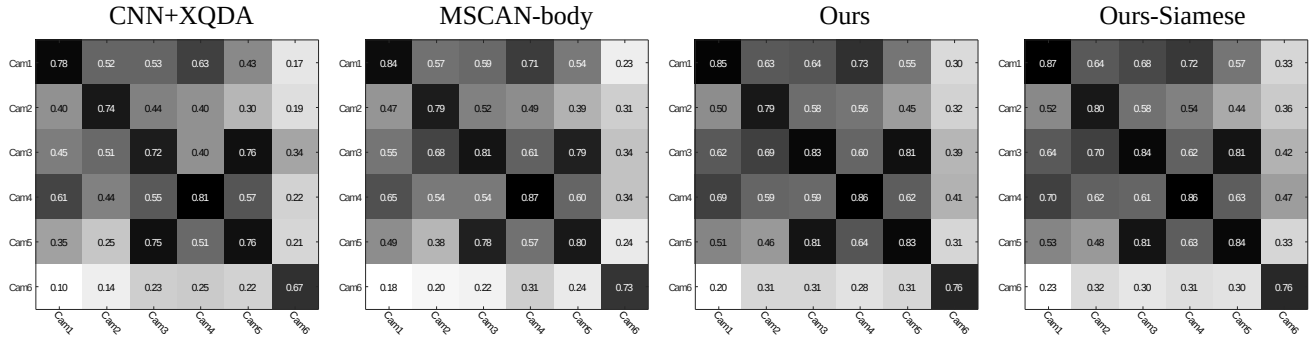

Figure 3. CMC rank-1 accuracy between camera pairs on MARS [5]. The compared methods include CNN+XQDA [5], MSCAN-body [2] with RGB-M as inputs, and our proposed MGCAM and MGCAM-Siamsese. Our MGCAM-Siamese performs the best.

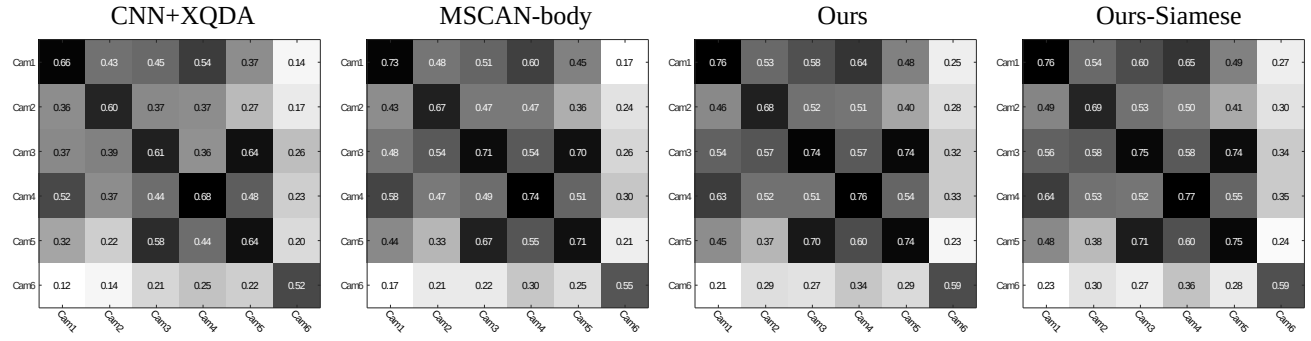

Figure 4. mAP accuracy between camera pairs on MARS [5]. The compared methods include CNN+XQDA [5], MSCAN-body [2] with RGB-M as inputs, and our proposed MGCAM and MGCAM-Siamsese. Our MGCAM-Siamese performs the best.

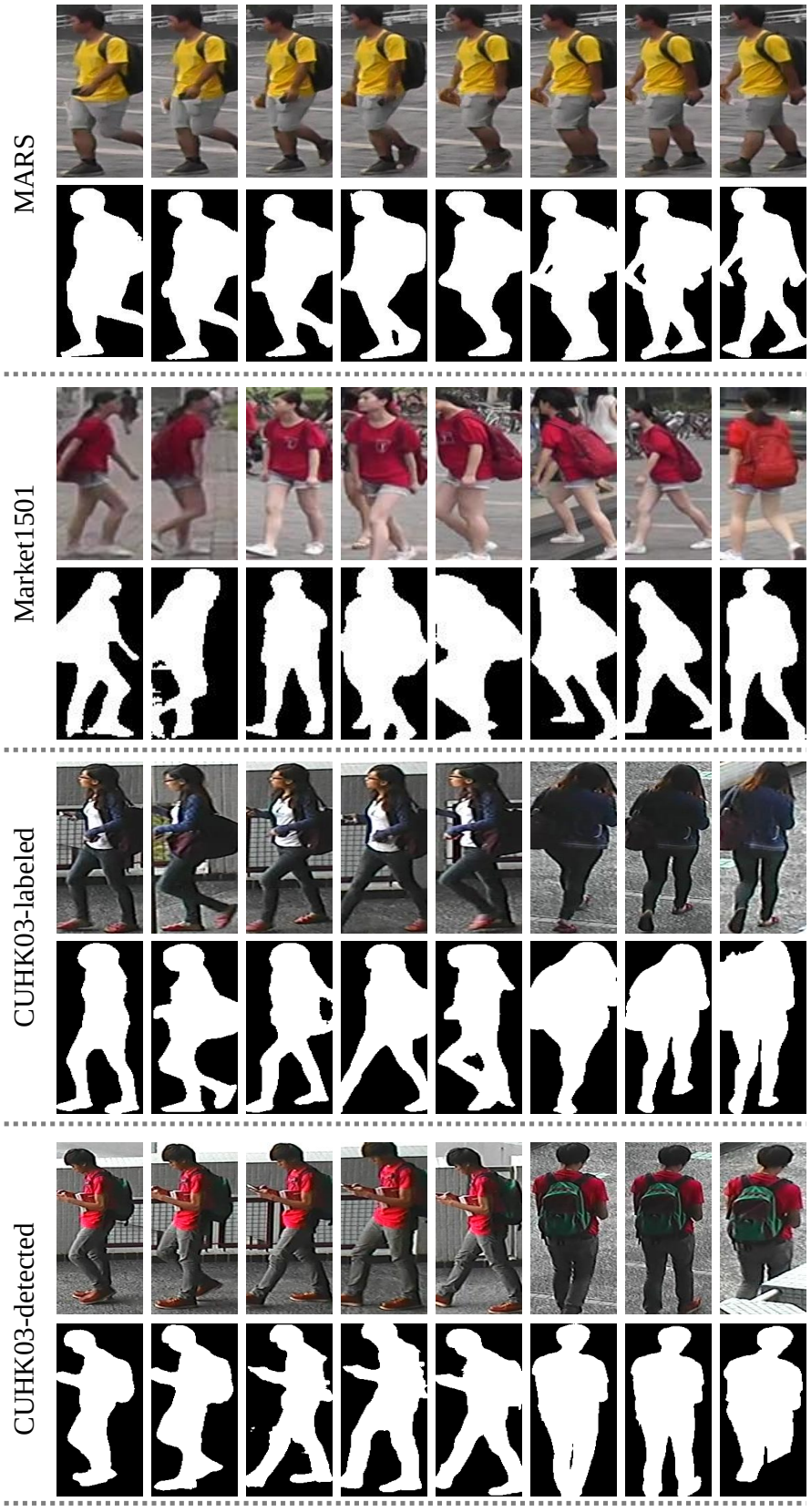

Figure 5. Segmentation examples from MARS [5], Market1501 [6] and CUHK03 [3]. Most masks are satisfying even for the images with complex backgrounds. There are also some failures caused by the wrongly detected images, i.e., person with bag and multiple persons in one image.