Supplementary of Multi-scale Deep Learning Architectures for Person Re-identification

Xuelin Qian[1] Yanwei Fu[2,5,*] Yu-Gang Jiang[1,3] Tao Xiang[4] Xiangyang Xue[1,2]
[1]Shanghai Key Lab of Intelligent Info. Processing, School of Computer Science, Fudan University;
[2]School of Data Science, Fudan University; [3]Tencent AI Lab;
[4]Queen Mary University of London; [5]University of Technology Sydney;
{15110240002,yanweifu,ygj,xyxue}@fudan.edu.cn; t.xiang@qmul.ac.uk

1. Multi-scale stream layers

Multi-scale-A layer (Fig. 1), analyses the data stream with the size 1×1, 3×3 and 5×5 of receptive field. Furthermore, in order to increase both depth and width of this layer, we split the filter size of 5×5 into two 3×3 streams cascaded (*i.e.* stream-4 and stream-3 in Tab 1 and Fig. 1). The weights of each stream are also tied with the corresponding stream in another branch. Such a design art is, in general, inspired by, and yet different from the inception architectures [11, 12, 10]. The key difference lies in the weights which are not tied between any two streams from the same branch, but are tied between the two corresponding streams of different branches.

Reduction layer (Fig. 2) further passes the data stream in multi-scale, and halves the width and height of feature maps, which should be, in principle, reduced from 78×28 to 39×14. We thus employ Reduction layer to *gradually* decrease the size of feature representations as illustrated in Tab 1 and Fig. 2, in order to avoid representation bottlenecks. Here we follow the design principle of "avoid representational bottlenecks" [12]. In contrast to directly use max-pooling layer for decreasing feature map size, our ablation study shows that the Reduction layer, if replaced by max-pooling layer, will leads to more than 10% absolute points lower than the reported results of Rank-1 accuracy on CUHK01 dataset. Again, the weights of each filter here are also tied for paired streams.

Multi-scale-B layer (Fig. 3) serves as the last stage of high-level features extraction for the multiple scales of 1×1, 3×3 and 5×5. Besides splitting the 5×5 stream into two 3×3 streams cascaded (*i.e.* stream-4 and stream-3 in Tab 1 and Fig. 3). We can further decompose the 3×3 C-filters into one 1×3 C-filter followed by 3×1 C-filter [10]. This leads to several benefits, including reducing the computation cost on 3×3 C-filters, further increasing the depth of this component, and prone to extract asymmetric features from the receptive field. We still tie the weights of each filter.

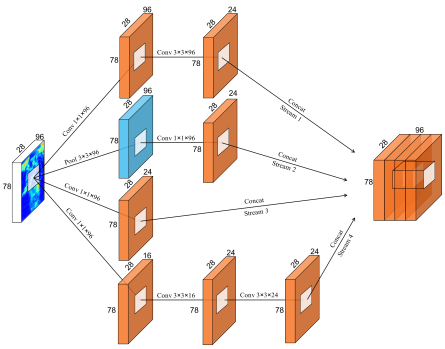


Figure 1. The structure of Multi-scale-A has four streams.

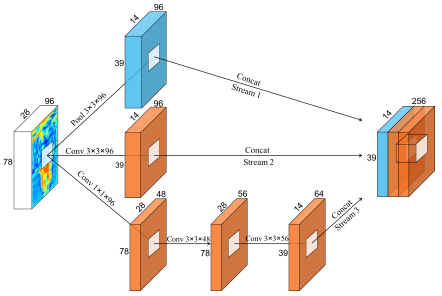


Figure 2. The structure of Reduction.

2. Experiment Results on CMC curves

We compare our proposed approach with the deep learning based methods including DeepReID [5], Imp-Deep [1], En-Deep [16], and G-Dropout [13], Gated_Sia [14], EMD [9], and SI-CI [15], as well as other non-deep competitors, such as Mid-Filter [18], and XQDA [8], LADF [7], eSDC [17], LMNN [4], and LDM [3], on three widely used datasets, CUHK01[6], CUHK03[5] and VIPeR[2] datasets.

*Corresponding Author

Layers	Stream id	number@size	output
Multi-scale-A	1	$1@3 \times 3 \times 96$ AF –$24@1 \times 1 \times 96$ CF	$78 \times 28 \times 96$
	2	$24@1 \times 1 \times 96$ CF	
	3	$16@1 \times 1 \times 96$ CF – $24@3 \times 3 \times 96$ CF	
	4	$16@1 \times 1 \times 96$ CF –$24@3 \times 3 \times 96$ CF– $24@3 \times 3 \times 24$ CF	
Reduction	1	$1@3 \times 3 \times 96$ MF*	$39 \times 14 \times 256$
	2	$96@3 \times 3 \times 96$ CF*	
	3	$48@1 \times 1 \times 96$ CF – $56@3 \times 3 \times 48$ CF– $64@3 \times 3 \times 56$ CF*	
Multi-scale-B	1	$256@1 \times 1 \times 256$ CF	$39 \times 14 \times 256$
	2	$64@1 \times 1 \times 256$ CF –$128@1 \times 3 \times 64$ CF–$256@3 \times 1 \times 128$ CF	$39 \times 14 \times 256$
	3	$64@1 \times 1 \times 256$ CF – $64@1 \times 3 \times 64$ CF –$128@3 \times 1 \times 64$CF– $128@1 \times 3 \times 128$ CF– $256@3 \times 1 \times 128$ CF	$39 \times 14 \times 256$
	4	$1@3 \times 3 \times 256$ AF* –$256@1 \times 1 \times 256$ CF	$39 \times 14 \times 256$

Table 1. The parameters of tied multi-scale stream layers of MuDeep. Note that (1) number@size indicates the number and the size of filters. (2) * means the stride of corresponding filters is 2; the stride of other filters is 1. We add 1 padding to the side of input data stream if the corresponding side of C-filters is 3. (3) CF, AF, MF indicate the C-filters, A-filters and M-filters respectively. A-filter is the average pooling filter.

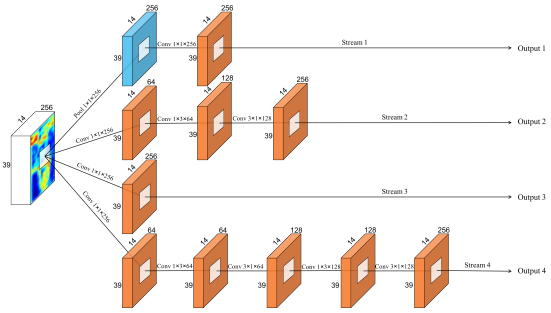


Figure 3. The structure of Multi-scale-B.

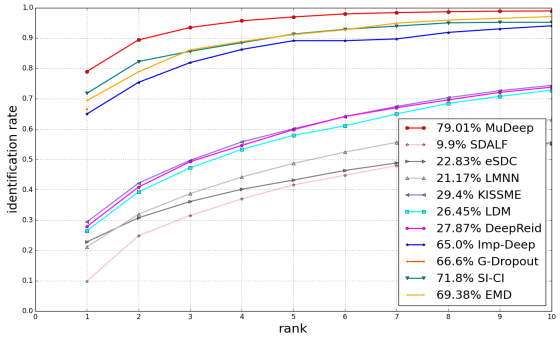


Figure 4. CMC curves of CHUK01 dataset.

All the Cumulative Matching Characteristics (CMC) curves are shown in Fig. 4, Fig. 5, Fig. 6, and Fig. 7.

For all the cases (both jointly and exclusively settings) of CUHK03 Detected/Labelled datasets, our MuDeep outperforms the other competitors by clear margins on all the rank accuracies; see a comparison in Fig. 6 and Fig. 7. Note that for the methods of Gated_Sia[14] and G-Dropout[13], the results of rank@K ($K > 10$) are unavailable; and thus those results did not draw in the figures.

For CUHK-01 dataset, the performance of the proposed methods is elucidated in Fig. 4. Our approach obtains 79.01% on Rank-1 accuracy, which can beat all the state-of-the-art; and is 7.21% higher than the second best method [15].

VIPeR dataset is extremely challenging due to small data size and low resolution. In particular, this dataset has relatively small number of distinct identities and thus the positive pairs for each identity are much less if compared with the other two datasets. However, our MuDeep still remains competitive and outperforms all compared methods, the results are illustrated in Fig. 5.

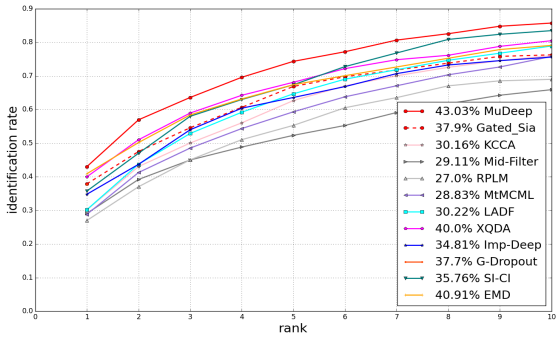
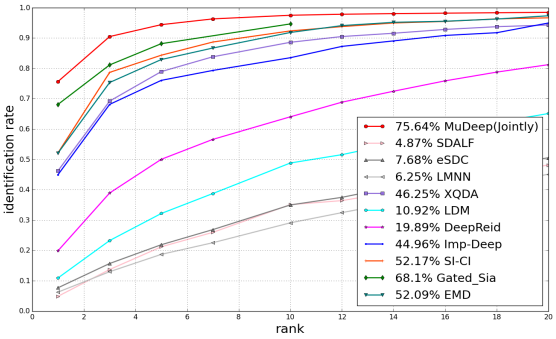


Figure 5. CMC curves of VIPeR dataset.

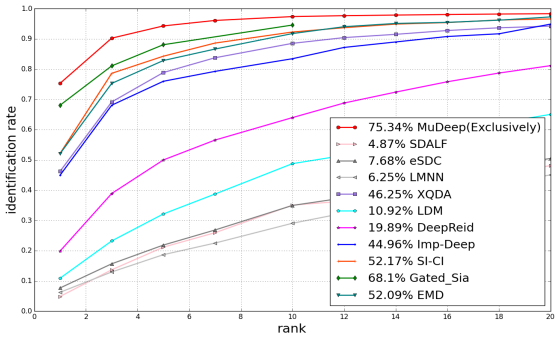
References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 2

[2] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE*

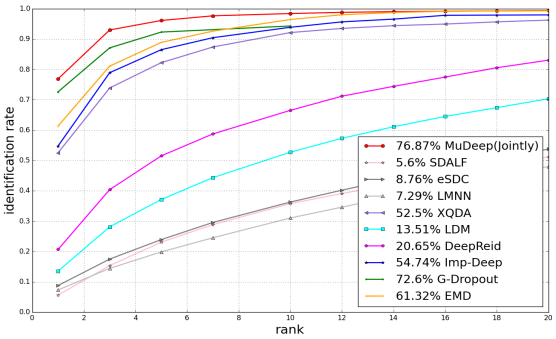


(a) Jointly setting of CHUK03-Detected

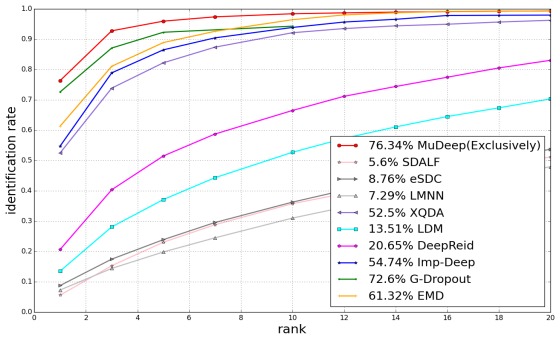


(b) Exclusively setting of CHUK03-Detected

Figure 6. CMC curves of CHUK03-Detected dataset. Mu-Deep is compared with the state-of-the-art.



(a) Jointly setting of CHUK03-Labelled



(b) Exclusively setting of CHUK03-Labelled

Figure 7. CMC curves of CHUK03-Labelled dataset. Mu-Deep is compared with the state-of-the-art.

PETS Workshop, 2007. 2

[3] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009. 2

[4] M. Hirzer, P. M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *IEE AVSS*, 2012. 2

[5] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[6] W. Li, R. Zhao, and X.Wang. Human re-identification with transferred metric learning. In *ACCV*, 2012. 2

[7] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *ECCV*, 2014. 2

[8] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2

[9] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei1, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 2

[10] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. In *arxiv*, 2016. 1

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Zbigniew-Wojna. Rethinking the inception architecture for computer vision. In *arxiv*, 2015. 1

[13] X. T, W.Ouyang, H. Li, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 2

[14] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 2

[15] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 2

[16] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *WACV*, 2016. 2

[17] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 2

[18] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 2