Bilinear Attention Networks for Person Retrieval Supplementary Material

Pengfei Fang[1,2], Jieming Zhou[1], Soumava Kumar Roy[1,2], Lars Petersson[1,2], Mehrtash Harandi[2,3]
[1]The Australian National University, [2]DATA61-CSIRO, Australia, [3]Monash University
{Pengfei.Fang, u5761794, Soumava.KumarRoy}@anu.edu.au
Lars.Petersson@data61.csiro.au, mehrtash.harandi@monash.edu

1. Ablation Study on Other Datasets

Here, we show an additional ablation study to verify the effectiveness of Bi-attention with AiA on the DukeMTMC-reID [3] and the MSMT17 [4] datasets in a single query setting. From this additional study, we can draw the same conclusions as in our main paper.

Effect of Bilinear Attention. We evaluate the effect of Bi-attention on the feature extractors, and the results are shown in Table 1. The results on both datasets demonstrate that: Bi-attention improves the performance of both scenarios $\mathcal{F}_a$ and $\mathcal{F}_a + \mathcal{F}_p$, similar to the observations noticed on the Market-1501 and CUHK03 datasets.

Table 1. Effect of Bi-attention on the DukeMTMC-reID [3] and MSMT17 [4] datasets.

		DukeMTMC-reID @ SQ		MSMT17 @ SQ	
Model		mAP	R-1	mAP	R-1
(i)	$\mathcal{F}_a$	69.0	83.2	39.3	65.8
(ii)	+ Bi-attention w/ AiA	71.0	84.4	47.4	72.7
(iii)	$\mathcal{F}_a + \mathcal{F}_p$	75.0	85.1	50.1	73.5
(iv)	BAT-net w/ AiA	**77.3**	**87.7**	**56.8**	**79.5**

Effect of the Position of Bilinear Attention. Table 2 shows the effect of Bi-attention when added to different positions of the baseline GoogLeNet network. It is clear that: adding Bi-attention in p_2 is superior compared to when it is added in p_1, p_3 and p_4. This verifies our conclusion that the feature maps in p_2 have richer channel information while still maintaining the spatial structural information, helping the network to focus more on the discriminative areas of the images.

Effect of the Dimensionality Reduction factor r. We further evaluate the effect of the reduction factor r in the embedding function $\varphi(\cdot)$. The results and comparisons shown in Table 3 reveal that: though r is an important parameter, affecting the size of the deep model, our network has only a weak dependency on r as changes in r lead to insignificant changes in the performance of our network on the DukeMTMC-reID and MSMT17 datasets.

Table 2. Effect of the position of Bi-attention on the DukeMTMC-reID [3] and MSMT17 [4] datasets.

		DukeMTMC-reID @ SQ		MSMT17 @ SQ	
Model		mAP	R-1	mAP	R-1
(i)	w/o attention	75.0	85.1	50.1	73.5
(ii)	p_1	76.4	86.8	53.8	77.0
(iii)	p_2	**77.3**	**87.7**	**56.8**	**79.5**
(iv)	p_3	75.8	86.3	50.3	74.2
(v)	p_4	75.2	85.8	50.1	73.7

Table 3. Effect of the dimensionality reduction factor r in the embedding function $\varphi(\cdot)$ on the DukeMTMC-reID [3] and MSMT17 [4] datasets.

		DukeMTMC-reID @ SQ		MSMT 17 @ D	
Model		mAP	R-1	mAP	R-1
(i)	w/o attention	75.0	85.1	50.1	73.5
(ii)	$r = 2$	77.2	87.5	55.9	78.8
(iii)	$r = 4$	**77.3**	**87.7**	**56.8**	**79.5**
(iv)	$r = 8$	77.2	87.6	56.4	78.7
(v)	$r = 16$	77.0	87.3	55.5	77.9
(vi)	$r = 32$	77.1	87.4	55.3	78.3

Visualisation of Bilinear Attention. We also visualise the Bi-attention for person images in both the DukeMTMC-reID dataset in Fig. 1(a) and the MSMT17 dataset in Fig. 1(b). Fig. 1 shows that: **(1)** the attention block masks out the non-informative background clutter in person images, and **(2)** the attention mask further emphasizes the discriminative parts of a person for the re-identification task, which reduces the prevalent misalignment problem in the retrieval task.

2. Further Analysis

Effect of Dimensionality of the Feature Embedding. The dimension, *i.e.*, Dim, of the feature embedding is evaluated and illustrated in Table 4 on the Market-1501 dataset [5] and the CUHK03 detected-set [2]. On Market-1501, we observe that mAP has a peak when $\text{Dim} = 768$ and for Rank-1 accuracy it peaks for $\text{Dim} = 512$. Thus we choose

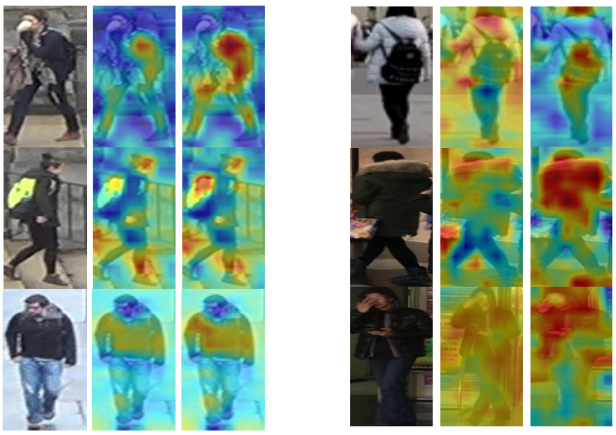

Figure 1. Visualisation of our Bi-attention in person images, sampled from the (a) DukeMTMC-reID dataset, and the (b) MSMT17 dataset. In each dataset, from left to right, (1) the input person image, (2) the input feature map to Bi-attention and (3) the masked feature map. In the heat map, the response increases from blue to red. Best viewed in color.

$\mathrm{Dim} = 512$ as it has an overall good performance and also reduces the amount of learnable parameters for the network. As for the CUHK03 dataset, it is clear that the feature embedding with $\mathrm{Dim} = 512$ performs well for both mAP and Rank-1 accuracy. Therefore, we choose $\mathrm{Dim} = 512$ as the dimension of the feature embedding across all datasets.

Table 4. Effect of the Dimensionality of Feature Embedding on the Market-1501 [5] and CUHK03 [2] datasets.

		Market @ SQ		CUHK03 @ D	
Model		mAP	R-1	mAP	R-1
(i)	$\mathrm{Dim} = 128$	85.9	94.3	71.0	74.1
(ii)	$\mathrm{Dim} = 256$	87.1	94.4	72.6	75.2
(iii)	$\mathrm{Dim} = 512$	87.4	**95.1**	**73.2**	**76.2**
(iv)	$\mathrm{Dim} = 768$	**87.6**	94.2	72.2	75.1
(v)	$\mathrm{Dim} = 1024$	82.9	93.5	72.3	74.9

Effect of Different Training Components. We further analyzed the effect of different training components (*e.g.*, random erasing, pre-training model) in Table 5 on the CUHK03 detected-set. Here, $\mathcal{F}_a$ and $\mathcal{F}_p$ denote the two parts of the network (see Fig.5 in the main paper). PRE and RE denote pretraining and random erasing, respectively. This table reveals that in all variations of those settings, adding our attention module leads to a significant boost in the mAP/R-1 values. Please note that in the aforementioned comparisons, all baseline performance is pre-trained and uses random erasing (*i.e.*, $\mathcal{F}_a + \mathcal{F}_p$ + PRE + RE).

Table 5. Effect of the Different Training Components on the CUHK03 [2] dataset. PRE and RE denote pretraining and random erasing, respectively.

		w/o Attention		w/ Attention	
Model		mAP	R-1	mAP	R-1
(i)	$\mathcal{F}_a + \mathcal{F}_p$	48.4	49.5	51.7	53.3
(ii)	$\mathcal{F}_a + \mathcal{F}_p$ + PRE	63.0	65.0	66.8	68.6
(iii)	$\mathcal{F}_a + \mathcal{F}_p$ + RE	52.1	53.5	58.9	61.6
(iv)	$\mathcal{F}_a + \mathcal{F}_p$ + PRE + RE	**67.8**	**71.1**	**73.2**	**76.2**

3. Discussion

Analysis of "Attention in Attention" and "Single Attention". In Table [1 - 4] of the main paper, we contrasted AiA against a simplified version, which still benefited from second order information (using bilinear pooling) but did not utilize the inner attention module (Fig. 4 vs. Fig. 1 in the main paper). Empirically, we observed that by incorporating the inner attention module, results could be improved. To further verify this, we replaced our AiA with a, so far, SOTA attention module, namely Squeeze and Excitation (SE) [1] and evaluated the resulting structure on the CUHK03 dataset with detected bounding boxes. The results are 70.3% / 72.8% (mAP/R-1) for SE and 73.2% / 76.2% for AiA, clearly showing the superiority of AiA.

Analysis of Failure Cases. In this section, we show some ranking lists of the failure cases across four datasets. Fig. 2 shows that the BAT-net may be affected by persons with similar distractors, such as similar clothing and stature. Further, for the case in DukeMTMC-reID (*i.e.*, the second ranking list), our system is also affected by occlusions (*i.e.*, car). Therefore, those limitations require us to develop a more robust person retrieval machine addressing these types of challenges.

References

[1] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[3] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[4] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person Trasfer GAN to Bridge Domain Gap for Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A

Figure 2. Some failure cases on the CUHK03 detected-set, Market-1501, DukeMTMC-reID and MSMT17 datasets from top row to bottom row. Here, the failure cases refer to the mismatching in Rank-1 retrieval. In each dataset, we list two cases. In each ranking list, to the left is the query person and to the right is the corresponding ranked list in the gallery set. The correct and false matches are enclosed in green and red boxes. Best viewed in color.

Benchmark. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.