Supplementary Material for "Explainable Person Re-identification with Attribute-guided Metric Distillation"

Xiaodong Chen[1]* Xinchen Liu[2] Wu Liu[2†] Xiao-Ping Zhang[3] Yongdong Zhang[1] Tao Mei[2]
[1]University of Science and Technology of China, Hefei, China
[2]JD AI Research, Beijing, China [3]Ryerson University, Toronto, Canada
cxd1230@mail.ustc.edu.cn,liuxinchen1@jd.com,liuwu@live.cn,xzhang@ee.ryerson.ca,zyd73@ustc.edu.cn,tmei@live.com

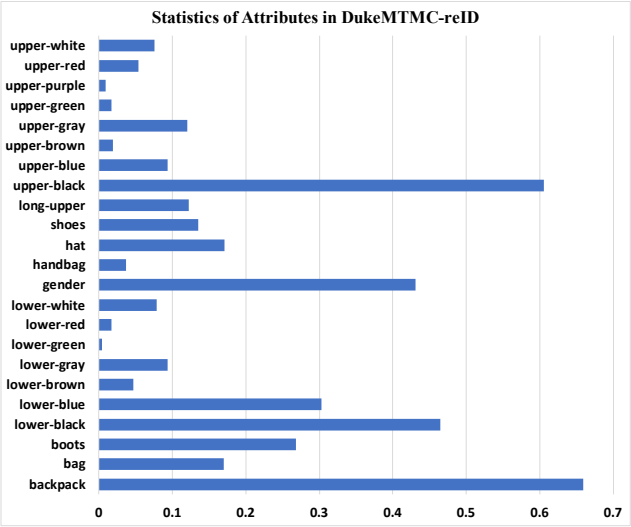

Figure 1. The attribute distribution of DukeMTMC-ReID [5].

In this supplementary material, we first introduce the attributes of the DukeMTMC-ReID [5] dataset. Then we provide additional experiments for 1) more visualizations of the learned attribute-guided attention maps (AAMs) on Market-1501 [4] and DukeMTMC-ReID [5] datasets, 2) visualizations of AAMs under the cross-domain setting, and 3) analysis on different designs of the interpreter networks.

1. Additional Experimental Results

1.1. Details of the DukeMTMC-ReID Dataset

For DukeMTMC-ReID, we select 23 attributes for the interpreter, i.e., gender (female/male), shoe type (boots/other shoes), wearing a hat (yes/no), carrying a backpack (yes/no), carrying handbag(yes/no), carrying other bags (yes/no), the color of shoes (dark/bright), length of upper clothing (long/short), 8 colors of upper clothing (black, white, red, purple, gray, blue, green, and brown) and 7 colors of lower clothing (black, white, red, gray, blue, green, and brown). The statistics of attributes on DukeMTMC-ReID [5] annotated by [2] are shown in Figure 1.

1.2. Visualizations of AAMs on Different Datasets

In this subsection, we exploit an interpreter learned for the target model, i.e., SBS (ResNet-50) [1] to further study the attention maps learned by the interpreter. Here we generate the average attention maps of individual attributes on the testing set of Market-1501 [4] and DukeMTMC-ReID [5], as shown in Figure 2 (a) and (b), respectively. In Figure 2 (a) or (b), the left part shows the positive average attention maps which are obtained from all images that contain a certain attribute. The right part shows the negative average attention maps obtained from all images that do not contain that attribute.

From the average attention maps, we can observe that: **1)** Overall we can find that the interpreter can effectively focus on the salient regions of most attributes, which is consistent with the observation of humans. **2)** For the large-area attributes such as the colors of upper and lower clothes, the interpreter can accurately focus on the corresponding regions. However, there are some worse examples, such as "low-green". This may be because these attributes have very few samples in the dataset, as shown in Figure 1. **3)** For those small but discriminative attributes like bag, hair, and handbag, the interpreter can also attend to the areas where the objects are most likely to appear. With this observation, we can figure out how the interpreter captures the differences between different attributes through attention.

Moreover, we have several interesting findings: **1)** The salient region of "gender" is mainly focused on the head and upper body of a person, which is similar to the attention of humans. 2) For DukeMTMC-ReID, the activation of "boots" is focused on both head and feet while the attention of "shoes" is only on the feet. This reflects that the ReID model learns biased knowledge about "boots" since

*This work was done when Xiaodong Chen was an intern at JD AI Research.
†Wu Liu is the corresponding author

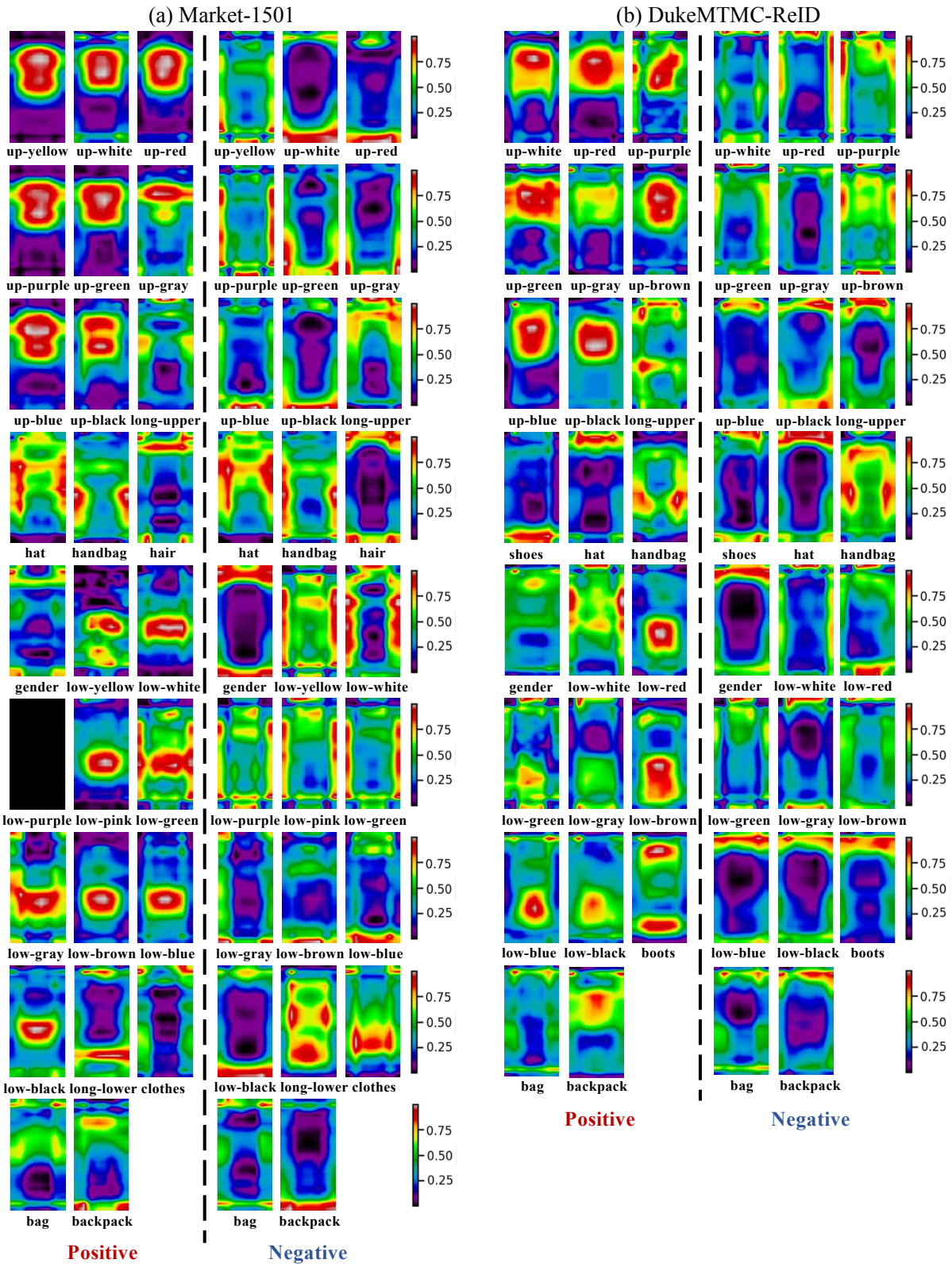

Figure 2. Visualization of average attention maps from the interpreter for the target model, SBS (ResNet-50) [1] trained on (a) Market-1501 [4] and (b) DukeMTMC-ReID [5]. In each sub-figure, the left part shows the average attention maps of each attribute which is obtained from all images that contain a certain attribute. The right part shows the average attention maps of each attribute obtained from all images that do not contain that attribute. (Best viewed in color.)

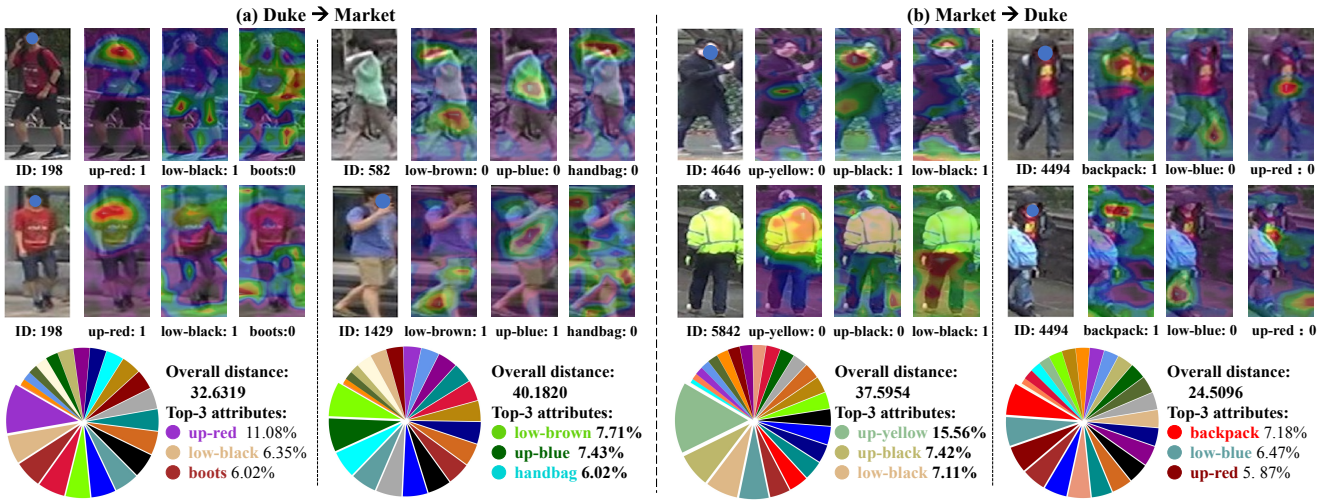

Figure 3. Pairwise examples and explanations for SBS (ResNet-50) under the cross-domain setting: (a) The interpreter is learned on DukeMTMC-ReID and applied to Market-1501 (Duke → Market); (b) The interpreter is trained on Market-1501 and tested on DukeMTMC-ReID (Market → Duke). For each pair of images, the upper part visualizes the AAMs of the top-3 attributes, which shows that the AAMs are attended to the discriminative attributes. The lower part provides the overall distance and contributions of the top-3 attributes to show the most contributed attributes discovered by the interpreter. (Best viewed in color.)

Model	N	Rank-1	mAP	X-mAP$_e$	X-mAP$_c$
ResNet-50	-	94.77	87.15	-	-
Interpreter	1	94.76	86.54	73.79	96.53
	2	93.98	86.27	74.31	96.36
	3	94.74	87.11	74.29	96.59
	4	94.52	86.91	74.47	96.79
	5	94.89	87.01	74.24	96.12

Table 1. Results of interpreters sharing different numbers N of stages from the target model. The results show that the interpreters have close performance for distance distillation.

the correlation among "boots", "hair", and "female" is very high as discussed in [2]. Therefore, the interpreter can help researchers and users find the biases in the datasets and improve the ReID models.

1.3. Interpretation for Cross-domain Setting

In Figure 3, we show several pairwise images and their explanations for SBS (ResNet-50) models trained on Market-1501 and DukeMTMC-ReID under the cross-domain setting. For each pair of images, we visualize the attribute-guided attention maps (AAMs) of the top-3 attributes and list the contributions of top-3 attributes to the overall distance.

From the AAMs, we can observe that for the attributes with similar distributions in two datasets, such as "up-red" and "backpack", the interpreter can effectively focus on corresponding regions under the cross-domain setting and make appropriate explanations. However, for the unique attributes of a certain dataset, e.g., "boots" that only exist

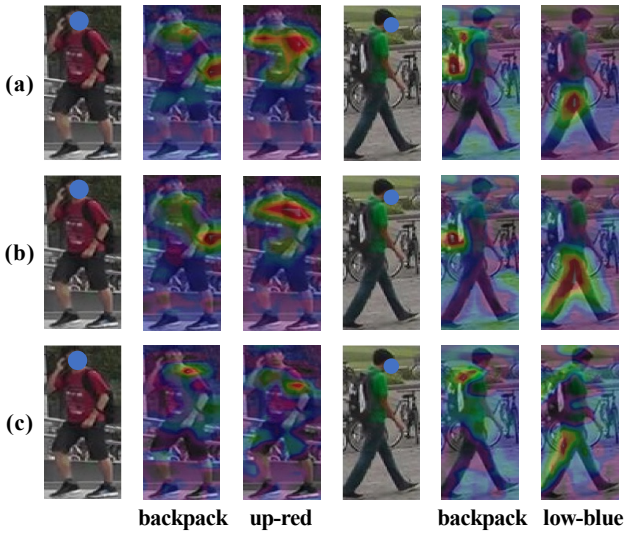

Figure 4. Attention maps from interpreters sharing different stages N from target models. (a) N = 1. (b) N = 3. (c) N = 5. The visualizations reflect the difference of interpreters using different levels of features from the target model. (Best viewed in color.)

in DukeMTMC-ReID, the interpreter would learn nothing about these attributes and generate incorrect explanations. To solve this problem, a straightforward strategy is to exploit a more comprehensive dataset with more diverse attributes, such as MTMS17 [3].

1.4. Study on Different Designs of Interpreter

We explore different designs of the interpreter by comparison of interpreter networks that share different numbers of stages from the target ReID model. Table 1 lists the re-

sults of interpreters sharing from only one stage to sharing all five stages ($N = 1 \sim 5$). We visualize several AAMs generated by different interpreters ($N = 1, 3, 5$) in Figure 4.

From Table 1, we can find that the quantitative metrics of all variants are very close. This means that all stages of the target ReID model may implicitly learn the knowledge to distinguish attributes. Only based on this observation, we might conclude that sharing five stages and only training the ADH module is the best choice since it needs to train much fewer parameters. However, as shown in Figure 4, the attention maps generated by different interpreters are of great difference. When $N = 1$, the attention maps are more scattered since the receptive field of the lower stage is relatively small which only focuses on local details. If $N = 5$, the attention maps are more focused on the centers of objects because the higher stages learn more high-level semantic concepts. By sharing parameters from the middle stage, i.e., $N = 3$, it can reach an equilibrium point between the low-level patterns and high-level semantics, which is more suitable for the representation of attributes. Therefore, our interpreter shares three stages from the target models.

References

[1] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *CoRR*, abs/2006.02631, 2020. 1, 2

[2] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 95:151–161, 2019. 1, 3

[3] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 3

[4] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2

[5] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *ICCV*, pages 3774–3782, 2017. 1, 2