ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language Supplementary Material

Zhe Wang[1]*, Zhiyuan Fang[2]*, Jun Wang[1], and Yezhou Yang[2]

[1] Beihang University
{wangzhewz,wangj203}@buaa.edu.cn
[2] Arizona State University
{zfang29,yz.yang}@asu.edu

This supplementary material includes implementation details of proposed methods and more qualitative results that are not presented in the main paper to demonstrate the performances of our Visual-Textual Attributes Alignment (ViTAA) model.

1 Implementation Details

Human parsing network. Our human parsing network is constructed based on HRNet [3]. The detailed information about the model is released in this project[1]. Due to the low resolution of pedestrian images captured in surveillance scenes, we slightly modify the network by changing the strides in the third transition layer to double the feature scale of the last resolution path. More parsing results are shown in Figure 1.

The network is first trained on the largest human parsing dataset MHPv2 [5], and then trained on ATR [1] which is one of the largest clothes parsing dataset. Lastly, the network is finetuned on the VIPeR [4] dataset. In this paper, the human parsing network works as a teacher network that distills the attribute knowledge to the lightweight segmentation layer in ViTTA though the generated attribute category annotations. We argue that these annotations could be also widely used on other tasks like re-id and pedestrian attribute analysis.

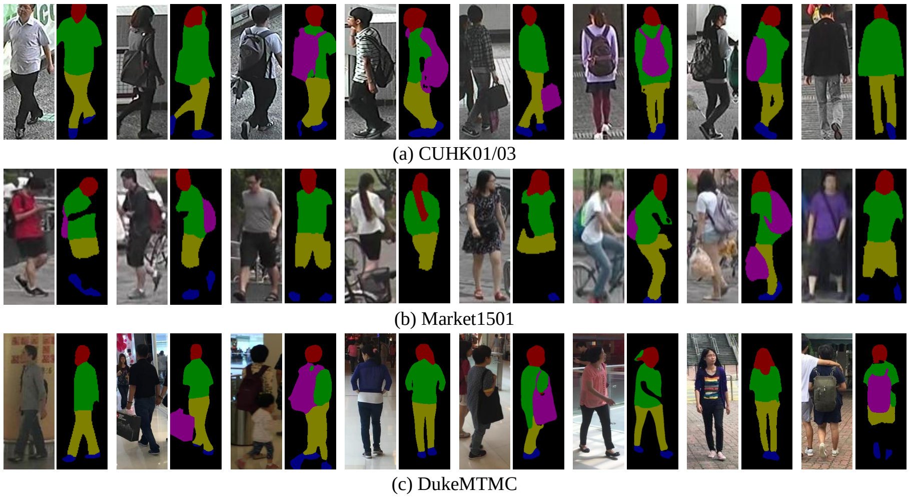
2 Experiments

More qualitative analysis. We provide more visual examination on the performance of ViTAA (as shown in Figure 2). Among all cases, ViTAA captures the attribute-phrases in queries and precisely locates them in the galleries.

Details of attribute retrieval. In order to validate our model's ability in visual attribute association, we conduct an extension experiment on attribute

* Equal contribution. This work was done when Z. Wang was a visiting scholar at Active Perception Group, Arizona State University.

[1] https://github.com/Jarr0d/Human-Parsing-Network



(a) CUHK01/03

(b) Market1501

(c) DukeMTMC

Fig. 1. Attribute annotation of different datasets generated by our human parsing network.

retrieval implemented on Market-1501 [6] and DukeMTMC [2]. In specific, we test only on the *upper-body clothing* attribute category, which is the attribute that existed in every data samples. To get the attribute retrieval accuracy, we label all the person images containing the target attribute with an identical ID for simple computation, then, we also remove the samples in queries that does not contain the target attribute during inference. We repeat this process for all the colors in *upper-body clothing* and report the retrieval results in Table 4 in the main paper.

References

1. Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., Lin, L., Yan, S.: Deep human parsing with active template regression. Pattern Analysis and Machine Intelligence, IEEE Transactions on (12), 2402–2414 (Dec 2015)
2. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision. pp. 17–35. Springer (2016)
3. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
4. Tan, Z., Yang, Y., Wan, J., Hang, H., Guo, G., Li, S.Z.: Attention-based pedestrian attribute analysis. IEEE transactions on image processing (12), 6126–6140 (2019)
5. Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., Feng, J.: Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: 2018 ACM Multimedia Conference on Multimedia Conference. pp. 792–800. ACM (2018)

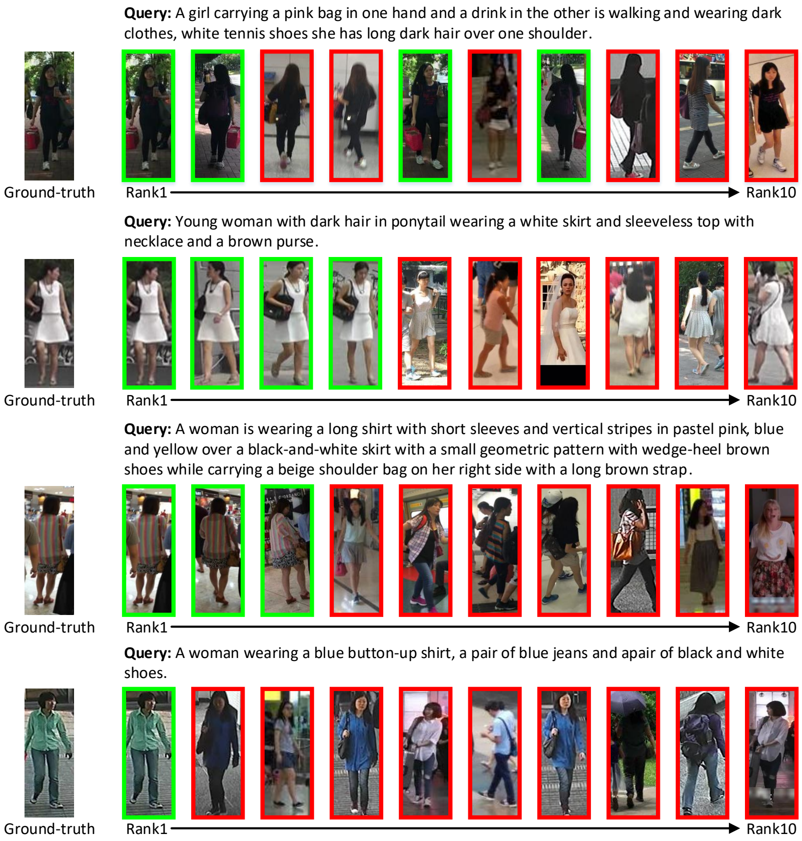

Query: A girl carrying a pink bag in one hand and a drink in the other is walking and wearing dark clothes, white tennis shoes she has long dark hair over one shoulder.
Ground-truth
Rank1
Rank10
Query: Young woman with dark hair in ponytail wearing a white skirt and sleeveless top with necklace and a brown purse.
Ground-truth
Rank1
Rank10
Query: A woman is wearing a long shirt with short sleeves and vertical stripes in pastel pink, blue and yellow over a black-and-white skirt with a small geometric pattern with wedge-heel brown shoes while carrying a beige shoulder bag on her right side with a long brown strap.
Ground-truth
Rank1
Rank10
Query: A woman wearing a blue button-up shirt, a pair of blue jeans and apair of black and white shoes.
Ground-truth
Rank1
Rank10

Fig. 2. Examples of person search results on CUHK-PEDES. We indicate the true/false matching results in green/red boxes.

6. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)