Co-Attention Aligned Mutual Cross-Attention for Cloth-Changing Person Re-Identification
Supplementary Material

Qizao Wang[1 [0000−0003−2556−5529]], Xuelin Qian[⋆2 [0000−0001−8049−7288]], Yanwei Fu[2 [0000−0002−6595−6893]], and Xiangyang Xue[1,2 [0000−0002−4897−9209]]

[1] School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
qzwang22@m.fudan.edu.cn, xyxue@fudan.edu.cn
[2] School of Data Science, and MOE Frontiers Center for Brain Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
{xlqian,yanweifu}@fudan.edu.cn

In the supplementary material, we provide more experiments on different designs of the Shape Semantics Embedding (SSE) module to show the effectiveness of our self-attention-based module design. We also provide more discussions and analyses about the t-SNE [4] visualizations shown in our main paper. Unless otherwise specified, the numbers of the figures and tables are within the scope of the supplementary material.

A. More Analysis of the SSE Module

To demonstrate the effectiveness and design rationality of our proposed Shape Semantics Embedding (SSE) module, we compare various module designs on the long-term cloth-changing person Re-ID dataset Celeb-reID [2].

Table 1. Ablation study of our proposed Shape Semantics Embedding (SSE) module on Celeb-reID. Details of each method are described and discussed in Sec. A.

Methods	Rank-1	mAP
Resnet-50 [1]	55.55	11.31
w/o Self	51.62	9.55
Self-v1	52.36	9.41
w/o SSE (Baseline)	52.86	9.92
Self-v2 (Ours)	**57.47**	**12.27**

Effectiveness of Self-Attention. As shown in Table 1, when we use the conventional benchmark network Resnet-50 [1] to extract body shape features from the heatmaps of human postures, the accuracy is inferior to "Self-v2", which is

⋆ corresponding author

our proposed SSE module. We argue that Resnet-50 is a strong CNN benchmark network, but it mainly focuses on small discriminative regions due to a Gaussian distribution of effective receptive fields [3]. However, the multi-head self-attention mechanism is good at capturing the long-distance and short-distance semantic information of inputs, so it is more effective to encode the rich correlations between human posture keypoints, which represent body shape semantic information. Additionally, compared with Resnet-50, a deep neural network composed of multiple convolution and pooling layers, our proposed SSE module is much more lightweight.

We also try discarding the self-attention module, which corresponds to "*w/o* Self" in Table 1. We observe a huge decrease in both Rank-1 and mAP, which shows the importance and effectiveness of the self-attention mechanism in capturing useful body shape semantic information to help handle the cloth-changing challenge.

Effectiveness of Our Self-Attention Design. If the encoding way is not designed well, body shape semantic information cannot be effectively gained even using self-attention. To prove it, we compare another SSE module design case, denoted as "Self-v1" in Table 1. It first directly applies an FC layer to encode the spatially flattened K heatmaps of human postures and gets $f^k \in \mathbb{R}^{K \times d}$ encoding keypoints information. Then apply one multi-head self-attention layer to capture correlations, followed by a convolution layer to adjust the output dimensions. As we can see, Rank-1 and mAP of "Self-v1" are much lower compared with our proposed method "Self-v2", and its mAP is even worse than the module design without self-attention, which demonstrates the rationality and effectiveness of our well-designed SSE module.

It is worth noting that the results of "*w/o* Self" and "Self-v1" are even worse than our baseline without the body shape branch. It demonstrates that if the body shape information is not correctly and effectively learned, the extra body shape branch is useless and even would affect the network to learn discriminative identity features by introducing noisy information. With our proposed SSE module, however, we can achieve the best results, which shows the effectiveness of our self-attention-based module design.

B. More Discussions of the t-SNE Visualizations

To verify our motivation and show the effectiveness of our proposed CAMC framework, we first randomly sample 4 pedestrians, with 3 clothes per person, and 10 images per clothing, on the LTCC [5] dataset. Then use t-SNE to visualize features from the appearance branch and the body shape branch, as well as ones extracted from CAMC.

As shown in Fig. 1, features from the appearance branch are relatively more chaotic than ones extracted from CAMC. As shown in Fig. 1 (a), some symbols are mixed, indicating different persons are misidentified under the influence of similar clothes. However, various symbols with the same color are clustered

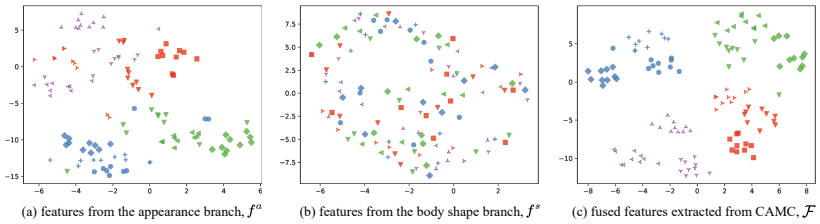

Fig. 1. Visualizations of features from the appearance branch and the body shape branch, as well as ones extracted from CAMC. Samples are randomly selected from the testing set of the LTCC dataset. Each color represents an identity, and different symbols indicate different clothes. Best viewed in color and zoomed in.

in Fig. 1 (c), which means the same person with various clothes is identified according to his/her cloth-irrelevant identity information.

It is worth noting that in Fig. 1 (b), not only symbols with the same color are not aggregated, but also symbols with the same shape distribute randomly in the feature space. It tells if being directly used, features from the body shape branch are not helpful to solve the cloth-changing problem of person Re-ID. It is understandable for the reason that we humans also cannot identify different pedestrians only from the heatmaps of their postures. However, we can distinguish people from RGB images based on their invariant biometric features, such as body shapes, even if they change clothes. Therefore, it is necessary to design a network to effectively interact between the appearance branch and the body shape branch, and distill useful identity information. As shown in Fig. 1 (c), our proposed CAMC framework can make use of body shape information to obtain discriminative fused features to tackle the cloth-changing problem of person Re-ID.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
2. Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P., Zhang, Z.: Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. IEEE Transactions on Circuits and Systems for Video Technology (2019)
3. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems **29** (2016)
4. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
5. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. In: Proceedings of the Asian Conference on Computer Vision (2020)