Supplementary Material for Conditional Cross Attention Network for Multi-Space Embedding without Entanglement in Only a SINGLE Network

Figure A1: Examples of Our TrainSets. The order of each row is FashionAI, DARN, DeepFasion and Zappos50K.

A. Datasets

FashionAI [6] The data published in the FashionAI Global Challenge 2018 has 180,335 apparel images. This comprises 8 fashion attributes containing 55 classes each.

DARN [1] An open dataset for attribute classification and street-to-shop image retrieval, comprising 253,983 images and 9 attributes. Each attribute contains 185 classes. The data is provided as image URLs; excluding broken URLs that cannot be downloaded, we used 195,771 URLs.

DeepFashion [2] This dataset comprises 289,222 images and 6 attributes. Each attribute contains 1000 classes.

Zappos50K [5] This dataset comprises 50,025 shoe images collected from Zappos.com. It consists of 4 attributes containing 34 classes each.

Algorithm 1 Pseudo-Code for CCA Training

```
input: Image $\mathcal{I}$, Condition $c$
batch $\mathcal{B}$, training epochs $K$, triplet set $\mathcal{T}$
Self Attention Block $SA$, Conditional Cross Attention $CCA$
for $epoch = 1, ..., K$ do
    for $\mathcal{B} = 1, ..., M \in \mathcal{T}$ do
        $Triplet(\mathcal{A}_c, \mathcal{P}_c, \mathcal{N}_c) \leftarrow \mathcal{B}$
        $\mathcal{I}, c \leftarrow \mathcal{A}_c, \mathcal{P}_c, \mathcal{N}_c$
        $\mathcal{Q}_i, \mathcal{K}_i, \mathcal{V}_i \leftarrow Token\_Embedding(\mathcal{I})$
        for $l = 1, ..., (\mathcal{L} - 1)$ do
            $\mathcal{Q}_i, \mathcal{K}_i, \mathcal{V}_i \leftarrow SA(\mathcal{Q}_i, \mathcal{K}_i, \mathcal{V}_i)$
        end for
        Last iteration $l = \mathcal{L}$ do
            $\mathcal{Q}_c \leftarrow Conditional\_Token\_Embedding(c)$
            $\leftarrow CCA(\mathcal{Q}_c, \mathcal{K}_i, \mathcal{V}_i)$
            $f \leftarrow l2(FC())$
        calculate $f_a, f_p, f_n \leftarrow Triplet(\mathcal{A}_c, \mathcal{P}_c, \mathcal{N}_c)$
        calculate triplet loss $\mathcal{L}(f_a, f_p, f_n|c)$
        calculate gradients of $\nabla\mathcal{L}(\theta)$
        $\theta \leftarrow Adam(\nabla\mathcal{L}(\theta))$
    end for
end for
```

Figure A1 presents actual examples using the four training sets. The figure shows four examples in the order of FashionAI [6], DARN [1], DeepFashion [2], Zappos50k [5].

B. More Visualization

B.1. Ranking and Attention Heat map

Figure A2 shows the Top 3 results along with each actual attention map. Each part of each attribute is considered, interpreted as the result of disentanglement multi-space modeling. The order in the figure is lapel design (notched), neckline design (round), skirt length (floor), pant length (midi), sleeve length (short), neck design (low turtle), coat length (midi), and collar design (peter pan).

B.2. Ours *vs*. Previous Works : Multi-Space Embedding

Figure 6 comparatively analyzed the embedding results of our study and previous studies. Of the 8 categories, the

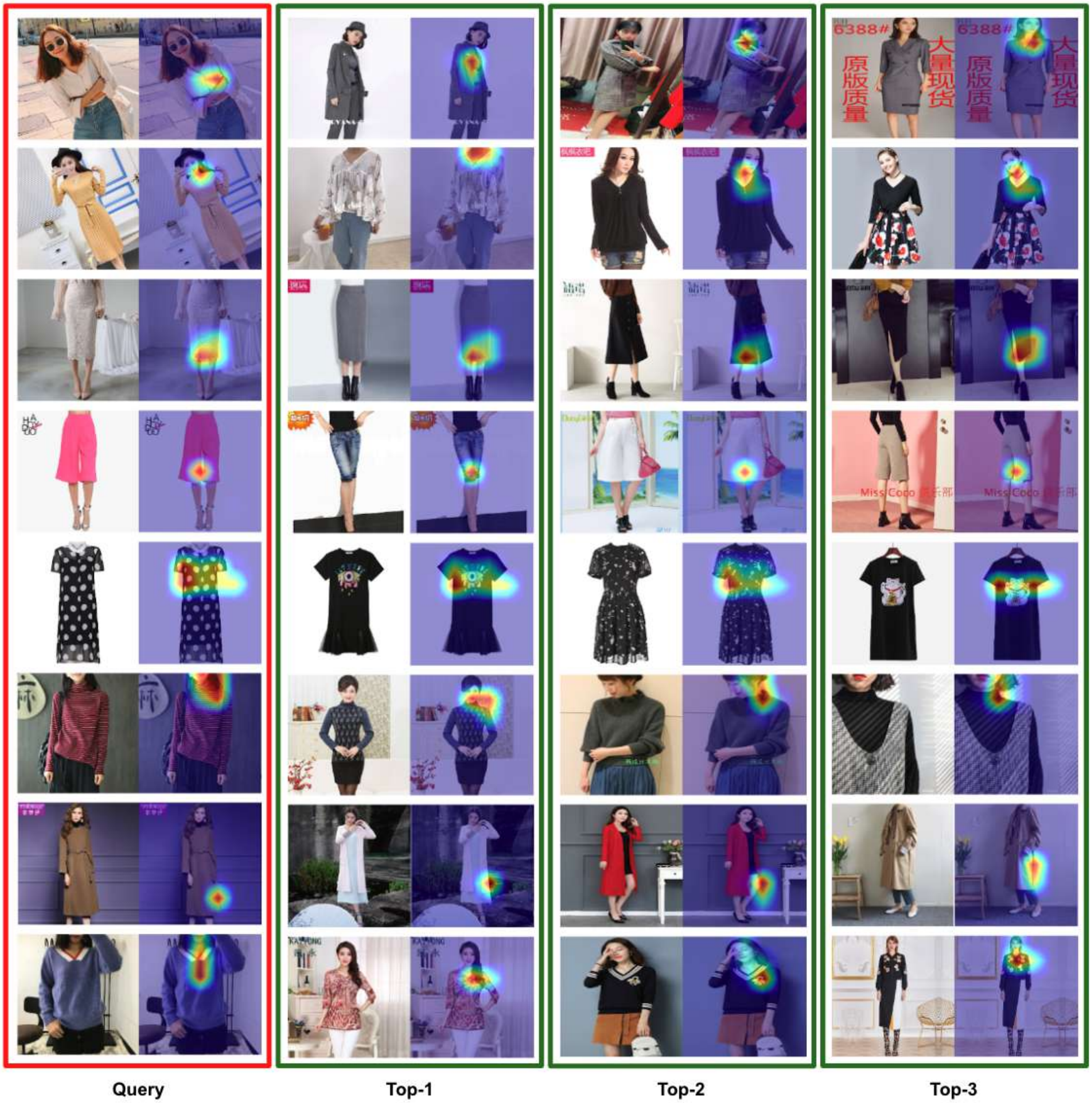
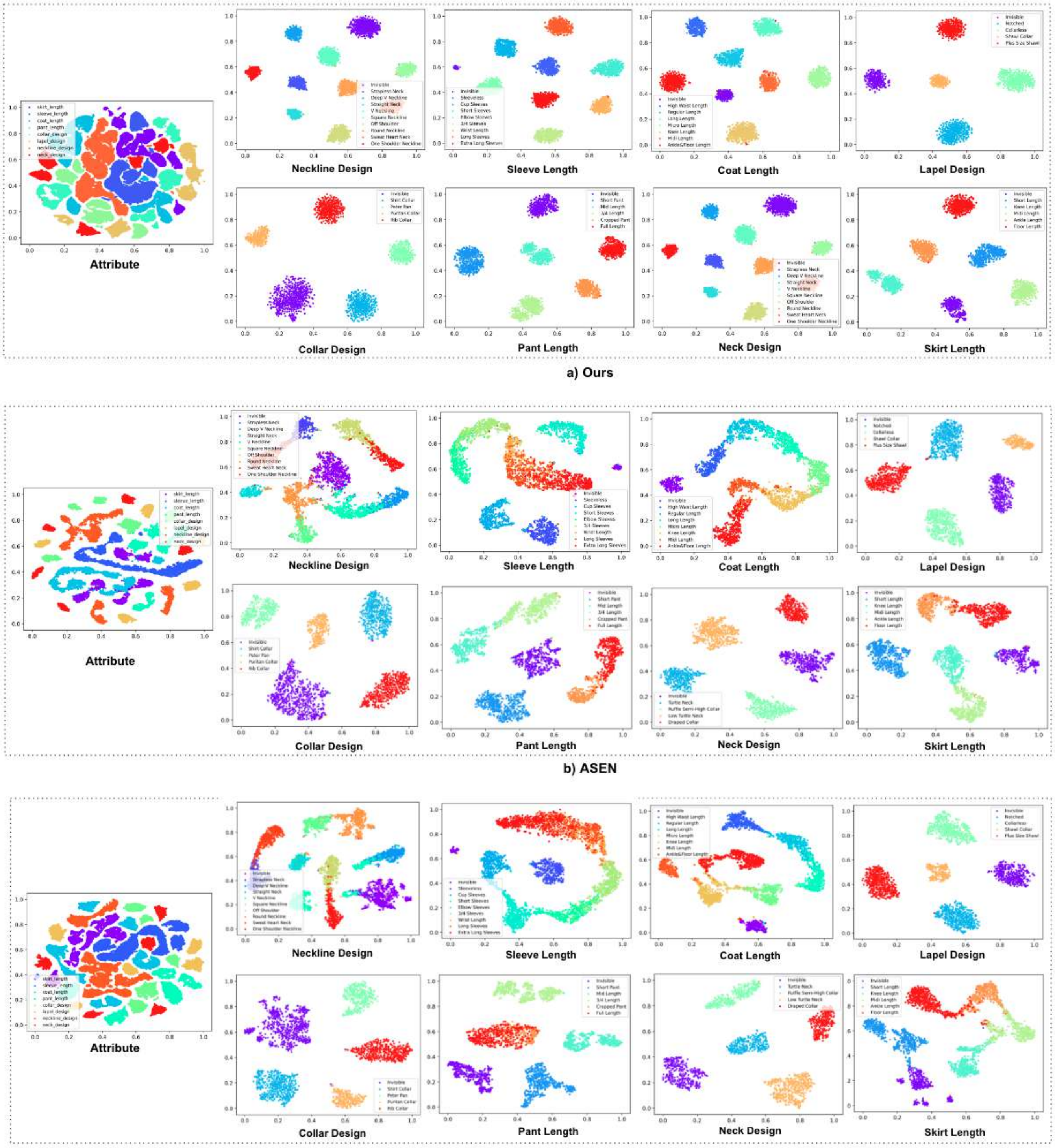


Figure A2: Examples of our top 3 ranking pair (image, attention heat map) results for FashionAI of 8 attributes. Red rectangle is query images. The order of each line is lapel design (notched), neckline design (round), skirt length (floor), pant length (midi), sleeve length (short), neck design (low turtle), coat length (midi) and collar design (peter pan).

results for Neck Design, Sleeve Length, and Coat Length were presented. Figure A3 shows the expanded results for all 8 categories. Our method solves the entanglement problem much better than ASEN [3] and CAMNet [4].

References

[1] Junshi Huang, Rogerio Feris, Qiang Chen, and Shuicheng Yan. Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network. In *ICCV*, 2015. 1



a) Ours

b) ASEN

c) CAMNet

Figure A3: Ours *vs*. Previous Works (ASEN, CAMNet) : Multi-space embedding's visualization using t-SNE about FashionAI.

[2] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*, 2016. 1

[3] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-Grained Fashion Similarity Learning by Attribute-Specific Embedding Network. In *Thirty-fourth AAAI Conference on Artificial Intelligence*, 2020. 2

[4] Chull Hwan Song and Hye Joo Han. Convolutional attribute mask with two-step attention for fashion image retrieval. In *26th International Conference on Pattern Recognition (ICPR), IEEE*, 2022. 2

[5] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 1

[6] Xingxing Zou, Xiangheng Kong, W. Wong, Congde Wang, Yuguang Liu, and Yuanpeng Cao. FashionAI: A Hierarchical Dataset for Fashion Understanding. In *CVPRW*, pages 296–304, 2019. 1