FiNet: Compatible and Diverse Fashion Image Inpainting Supplemental Material

Xintong Han[1,2] Zuxuan Wu[3] Weilin Huang[1,2] Matthew R. Scott[1,2] Larry S. Davis[3]
[1]Malong Technologies, Shenzhen, China
[2]Shenzhen Malong Artificial Intelligence Research Center, Shenzhen, China
[3]University of Maryland, College Park
{xinhan,whuang,mscott}@malong.com, {zxwu,lsd}@umiacs.umd.edu

Network Details

We illustrate the detailed network structures of our shape generation network and appearance generation network in Figure 1 and Figure 2, respectively. There are some details to be noted:

• Our residual block module R^2 contains two residual blocks [4] with one 1×1 convolution in the beginning to make the number of input and output channels consistent after concatenation with the latent vector. We use 3×3 convolution for all the other convolutional operations in our network.

• We use the softmax function at the end of our shape generation network to generate segmentation maps, and the *tanh* activation is applied when our appearance generation network outputs synthesized images. Other activations utilize *ReLu*. No batch normalization is used in our network.

• To create layouts / images with a missing fashion item (*i.e.*, shape context $\hat{S}$ and appearance context $\hat{I}$), we need to mask out pixels of a fashion item, which is determined by the plausible region of a specific garment category. Given the human parsing results generated by [2], for a top item, we mask out the bounding box covering the regions of the top and upper body skin; for a bottom item, its plausible region contains bottom and lower body skin; for both hats and shoes, we use the bounding box of the corresponding fashion item to decide which region to mask out. The bounding boxes are slightly enlarged to ensure full coverage of these regions.

• For both shape and appearance, the inpainted region is first resized to 256×256, and the reconstruction losses are only computed over this resized inpainted region. Finally, we paste this region back to the input image and obtain the final result.

• For constructing contextual garments, we first utilize human parsing results S, generated by [2], to extract an image segment for each garment in its corresponding plausible region. We refer an *image segment* to a cropped RGB region with S from I—the image segments are extracted using S (by cropping $S \odot I$). The inputs to E_{cs} and E_{ca} are the same as in Figure 3 and 4 in the main paper. We tried to only input binary maps to the shape generation network but found that it yields worse results than with RGB image segments. This is because RGB image segments contain richer and more accurate compatibility information than binary maps. For example, even though man's shoes and women's shoes have very similar binary segmentation, they should have different impacts on the generated shape for a bottom.

Then, we resize these extracted image segments to 128×128, and concatenate them in the order of *hat*, *top*, *bottom*, *shoes* with the target garment category one set to all 1's (*e.g.*, *top* is the target category in Figure 1 and 2). This not only encodes the information of all contextual garments but also tells the network which category is missing. Finally, we have a $128 \times 128 \times 12$ contextual garment representation x_c.

• To further guarantee fair comparison between FiNet and others [10, 1, 6, 7], we modify the main network structures for all these methods to be the same as ours instead of using their original ones, which usually have worse performance than U-Net with residual blocks that we use as shown in Figure 1 and 2.

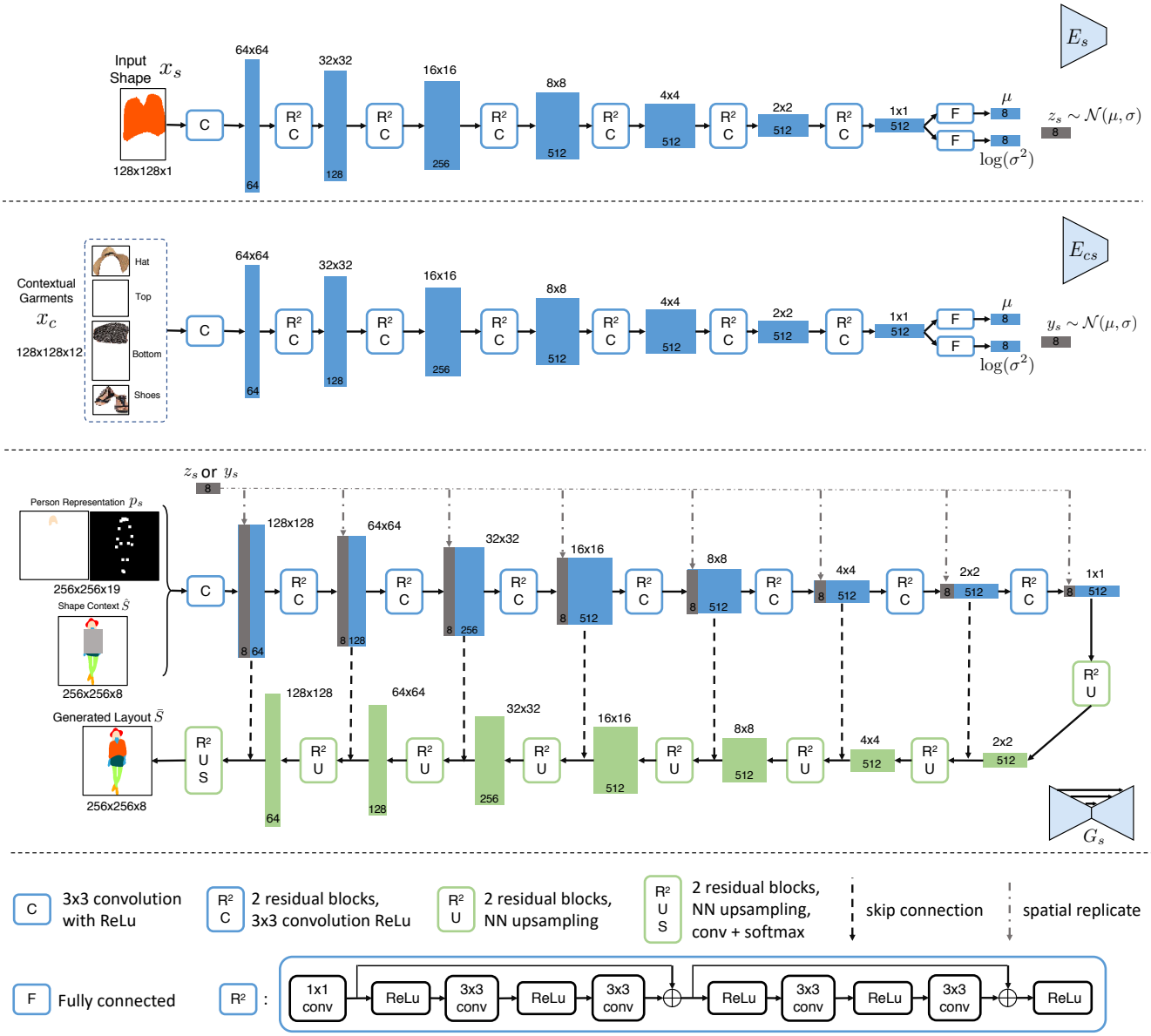

Figure 1: Network structure of our shape generation network.

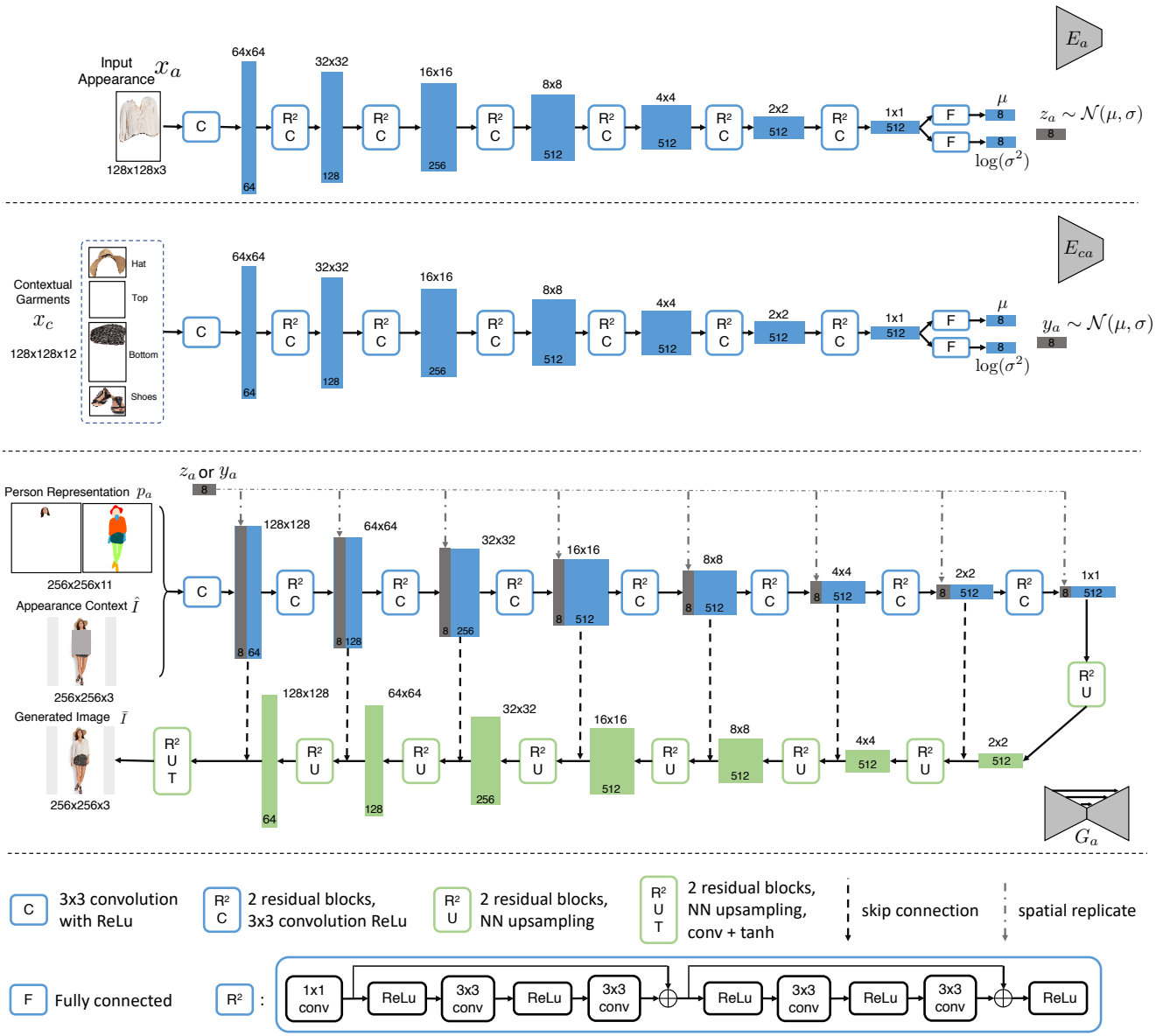

Figure 2: Network structure of our appearance generation network.

Latent Space Visualization

Shape latent space. In Figure 3, we visualize the generated segmentation maps of an image when varying different dimensions of the shape latent vector z_s. In the left example, the shape generation network generates different top garment layouts when we change the values of the 5-th, 6-th, and 7-th dimension of z_s (note that they are also the corresponding dimensions of y_s since y_s and z_s share the same latent space). We can find that different dimension controls different characteristics of the generated layouts: the 5-th dimension mainly controls the sleeve length—long sleeve $\rightarrow$ middle sleeve $\rightarrow$ short sleeve $\rightarrow$ sleeveless; the 6-th dimension determines the length of the clothing as well as the sleeve; and the 7-th dimension measures if the top opens in the middle. As for the right example, in which we generate bottom garment, the 6-th dimension is related to how the bottom interacts with the top; the length of the pants are changed when we vary the 7-th dimension; and the last dimension correlates with the exposure of the knee. Note that, for different garment categories, the same dimension of z_s (or y_s) controls different characteristics. For example, $z_{s,7}$ has something to do with the length of bottoms but not the length of tops.

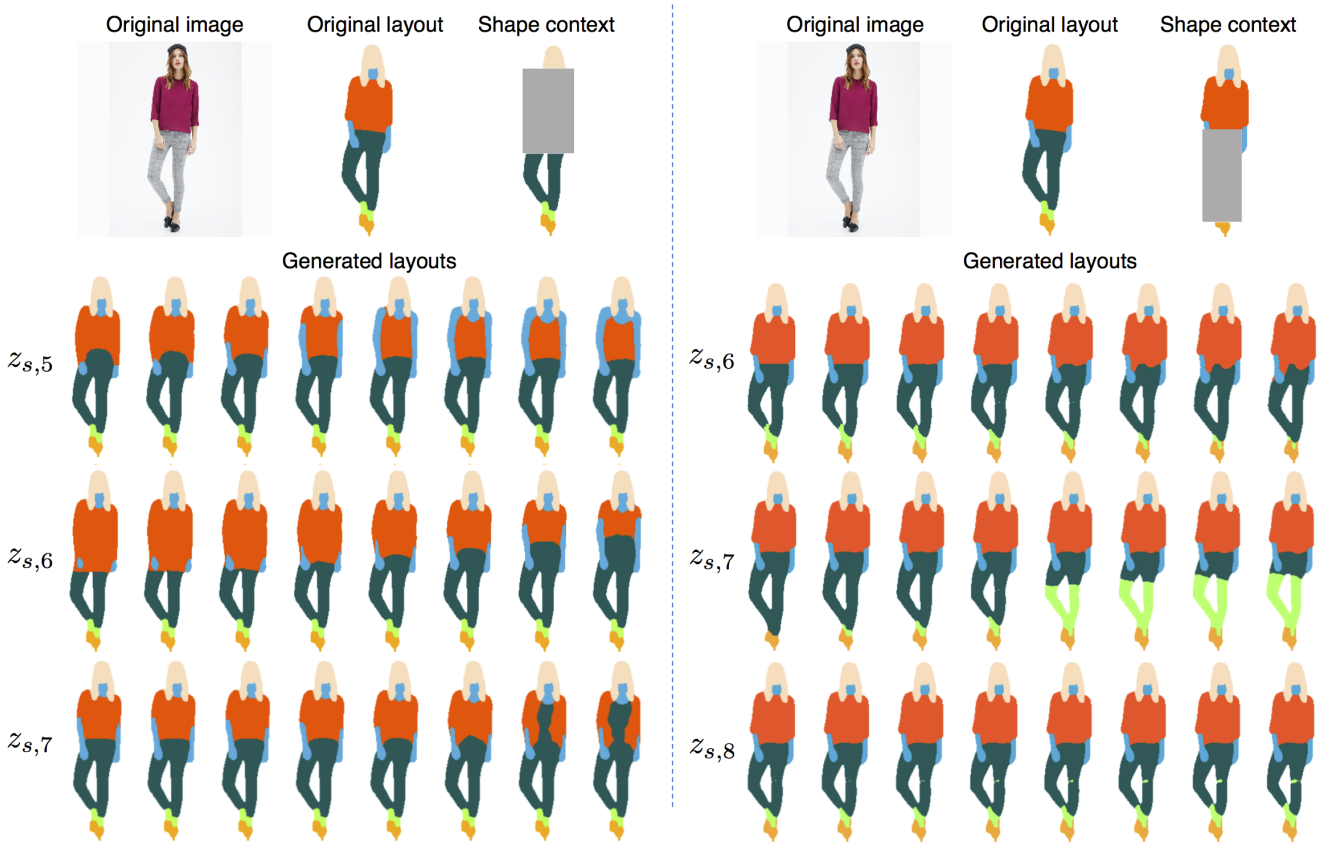

Figure 3: Generated layouts by our shape generation network when we change the values in different dimensions of the learned latent vector.

Further, in Figure 4 and 5, we show the compatibility space for two images by projecting the generated layouts in a 2D plane, whose x and y axes correspond to the 5-th (sleeve length) and 6-th (clothing length) dimension of y_s that is used to generate these layouts. μ_i and σ_i represent the mean and standard deviation of y_s's distribution in its i-th dimension. Consequently, layouts corresponding to latent vectors that are far from μ are uncompatible and unlikely to be generated.

In Figure 4, as we want to generate a compatible top for a man wearing a pair of long pants, the generated top layouts usually have long or short sleeves; and sleeveless tops (images in the lower right corner) are less compatible and realistic, thus these layouts are not likely to be generated (outside of 3σ). In contrast, when we generate layouts for a girl with a pair of shorts as shown in Figure 5, the generated layouts tend to have shorter sleeve length as well as clothing length because they are more compatible with shorts. By the comparison between Figure 4 and 5, we can see that our shape generation network

effectively models the compatibility and can generate compatible garment shapes according to contextual information.

There are some unrealistic sleeve shapes presented in Figure 4. Due to the continuous nature of Normal distribution, one cannot guarantee that every code in the latent space corresponds to a realistic shape. We tried to add adversarial training to the shape generation network, but found it is still hard to rule out all unrealistic cases. A potential solution is to have a discrete latent space with finite samples, which, however, will need more complicated modeling (*e.g.*, [5]). This is beyond the scope of this paper, and we consider it as a future research direction. Note that many of these unrealistic shapes fall out of 3σ. This indicates that unrealistic shapes are often incompatible and can be avoided to some extent by our shape generation network.

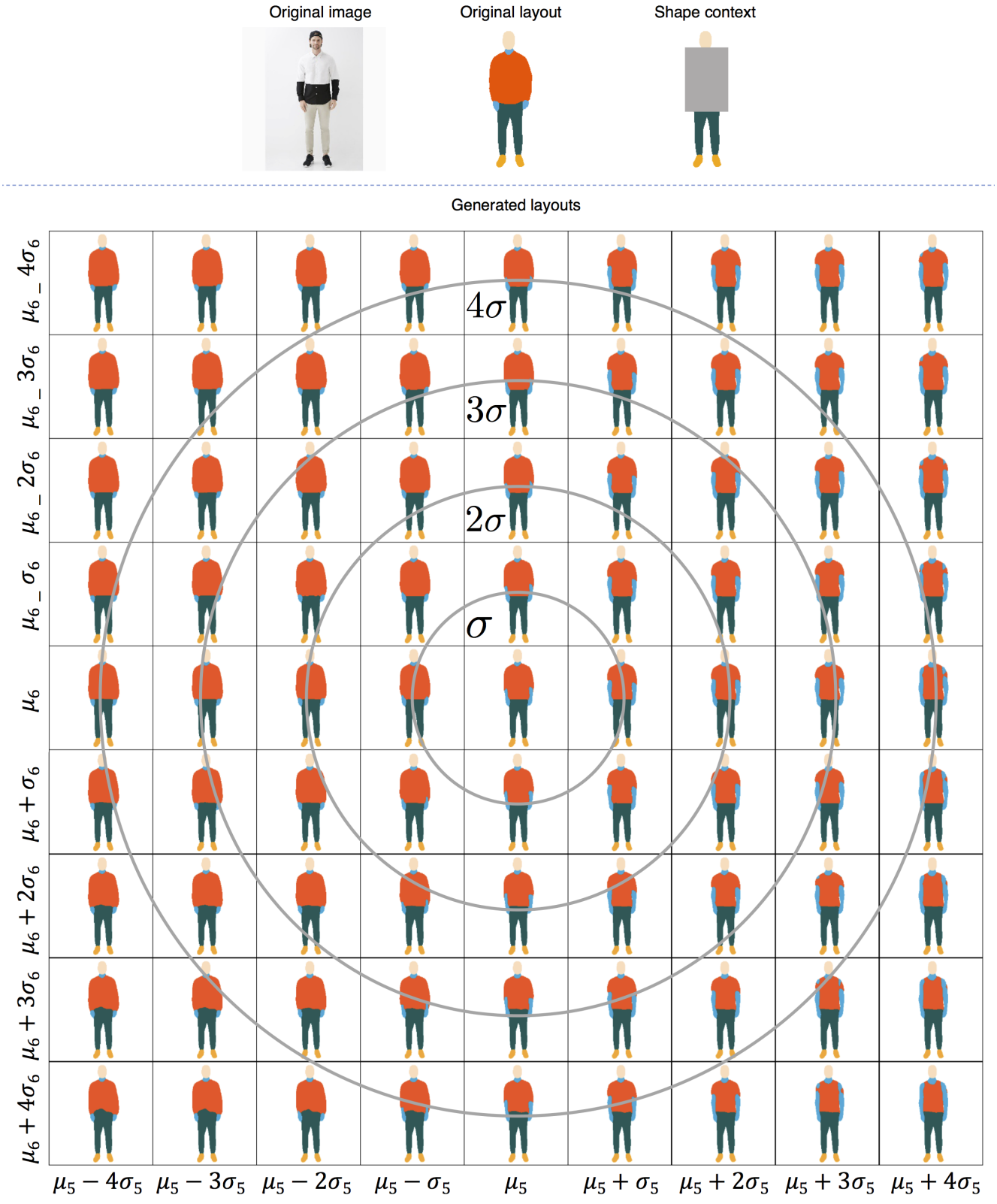

Figure 4: Shape compatibility space visualization. x and y axes correspond to the 5-th and 6-th dimensions of y_s, respectively.

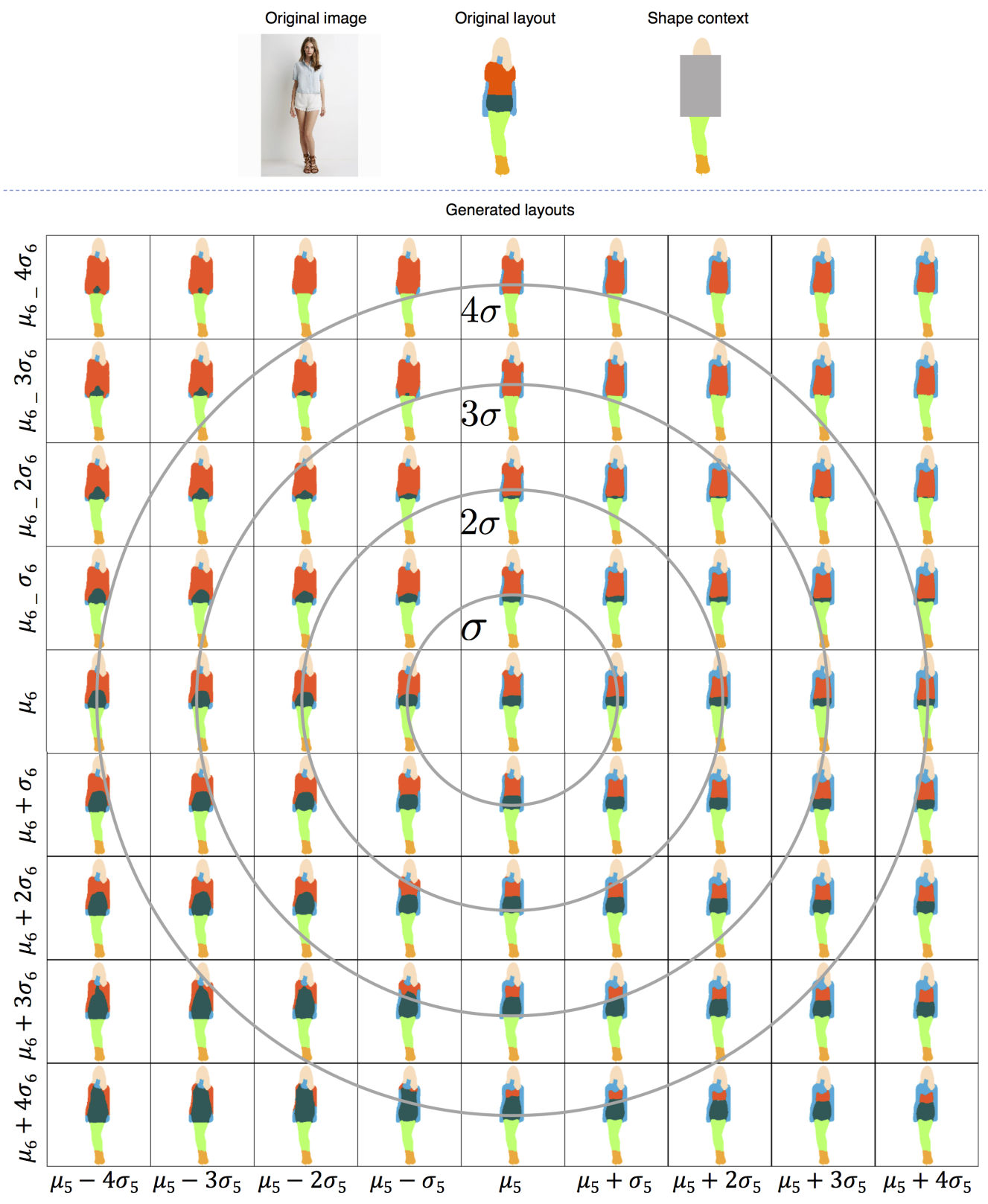

Figure 5: Shape compatibility space visualization. x and y axes correspond to the 5-th and 6-th dimensions of y_s, respectively.

Appearance latent space. As shown in Figure 6, we further present similar visualization for the appearance latent vector z_a (or y_a). Note that for visualization purposes, we use the ground truth segmentation map to generate appearance for simplicity. Unlike shape, the same dimension of the appearance latent vector correlates to the same appearance characteristic for different garment categories. The 1-st, 3-rd, 5-th dimensions correspond to brightness, color, texture of the generated images, respectively. This also indicates the importance of learning a compatible space; otherwise, if we project all appearances in the same latent space as FiNet w/o comp or BicyleGAN [10] without conditioning on the contextual garments, incompatible and visually unappealing shoes (shoes of all different colors as in the right side of Figure 6) may be generated.

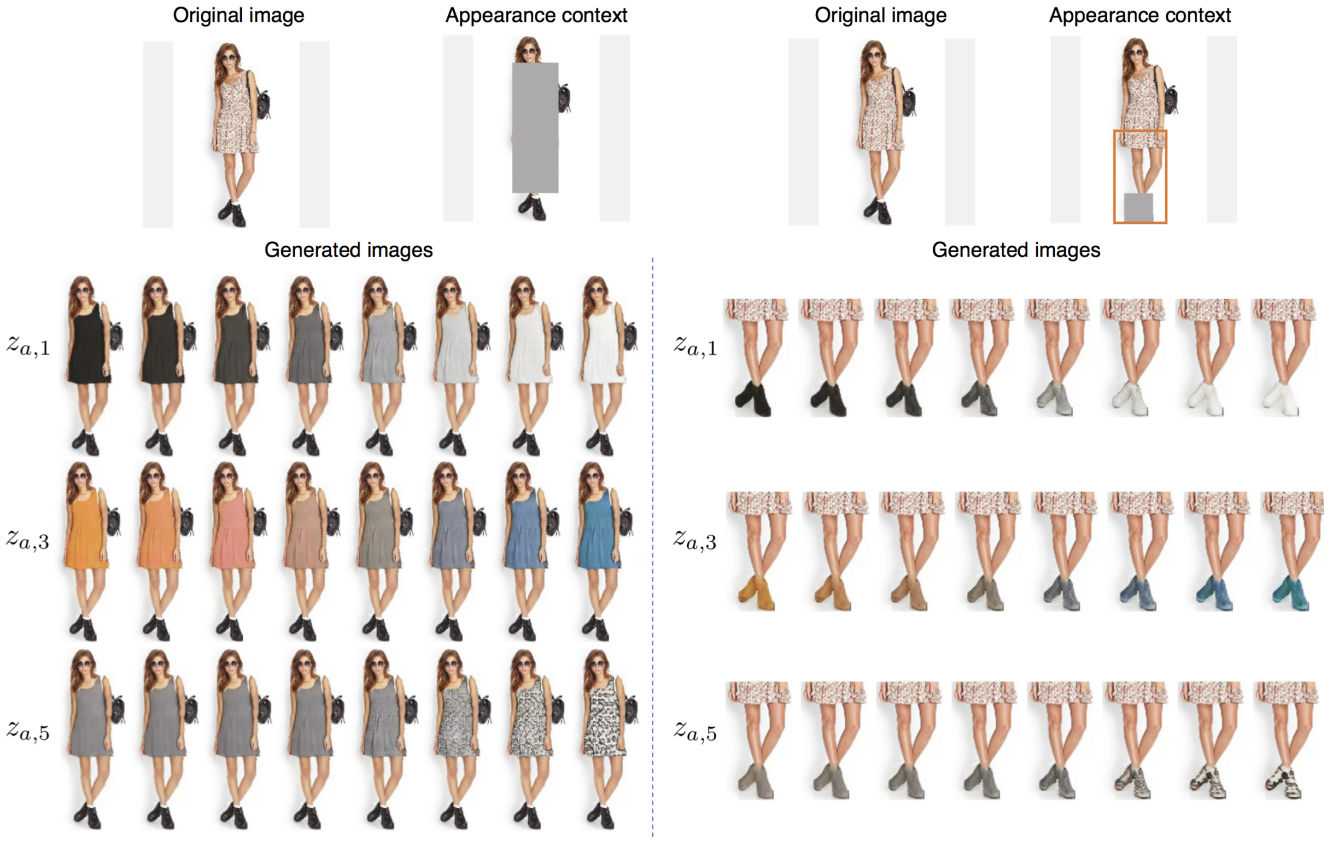

Figure 6: Generated layouts by our appearance generation network when we change the values in different dimensions of the learned latent vector.

We further plot the appearance compatibility space for three exemplar images in Figure 7, 8 and 9 for better understanding our appearance generation network. The generated appearances are arranged according to the 1-st (brightness) and 3-rd (color) dimension of y_a. In particular, in Figure 7, given the gray top, our network considers dark bottoms of black or blue as compatible, and the ground truth pants also present these visual characteristics. In Figure 8, we can find that since the white graphic T-shirt is more compatible with lighter bottoms, our generation network creates such shorts accordingly. Unlike these two cases, in Figure 9, there is no strong constraint in the color of a compatible dress, so the appearance generation network outputs dresses with various colors. The results illustrated in these figures again validate that we inject compatibility information into our network to ensure diverse and compatible image inpainting results.

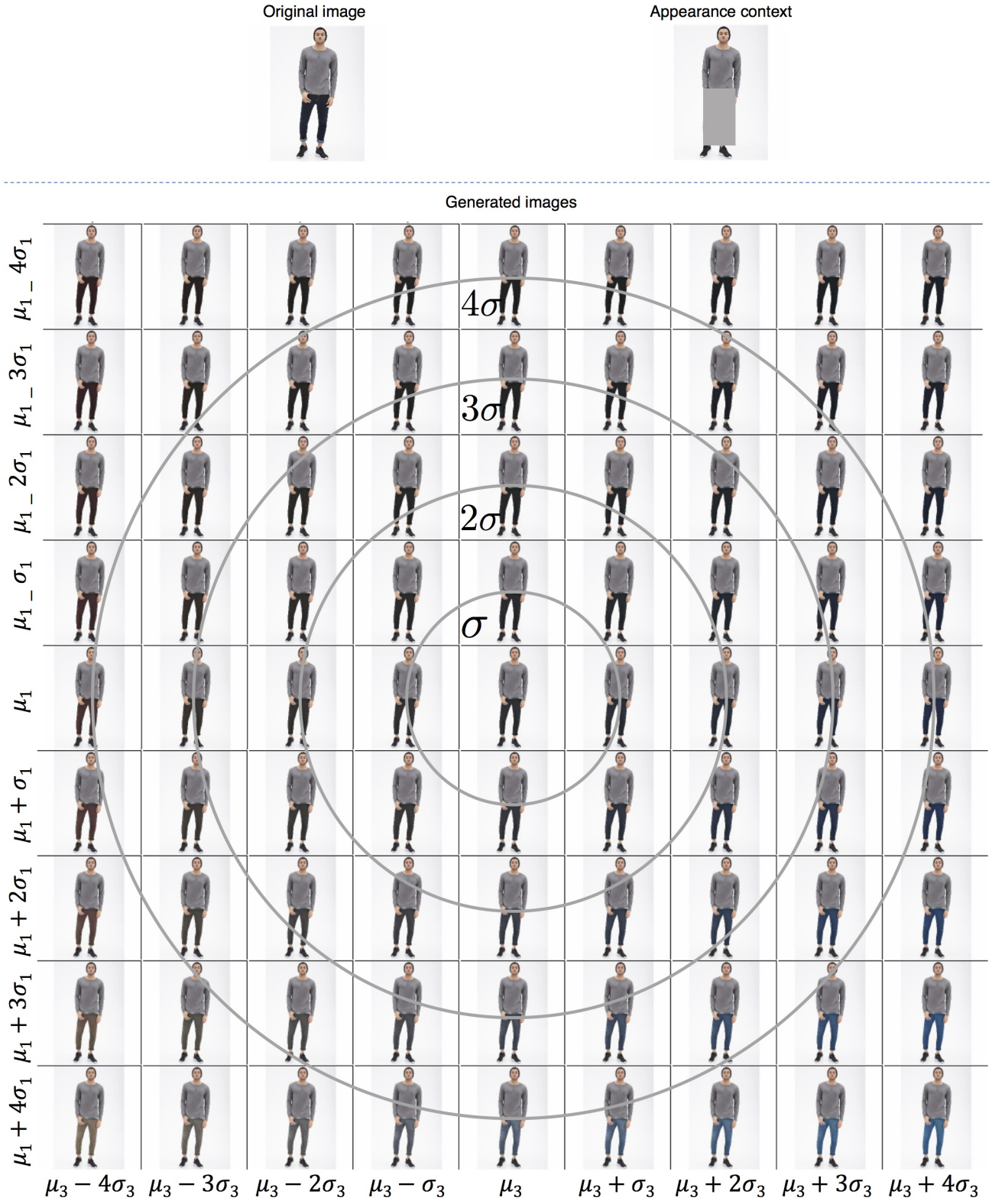

Figure 7: Appearance compatibility space visualization. x and y axes correspond to the 3-rd and 1-st dimensions of y_a, respectively.

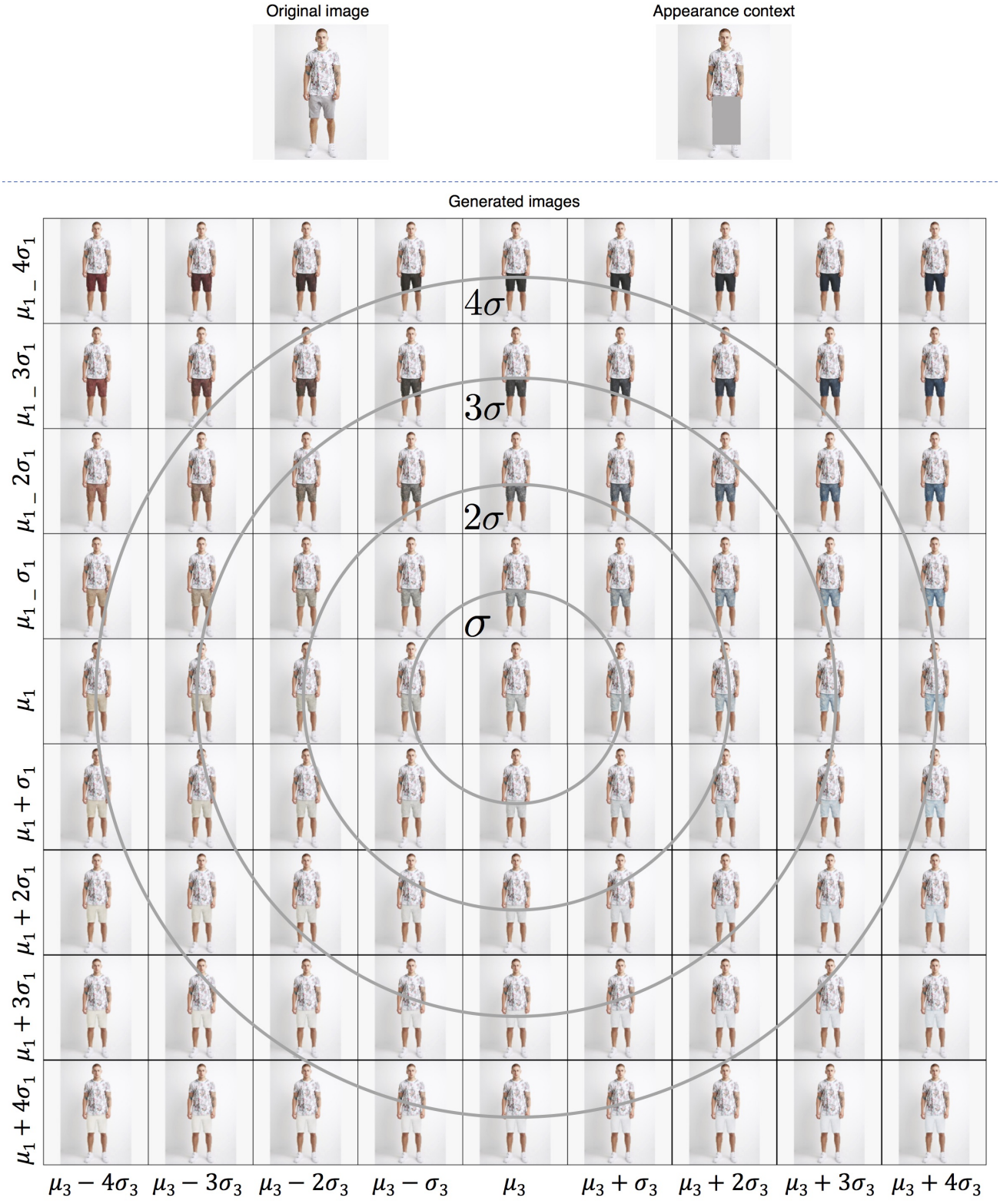

Figure 8: Appearance compatibility space visualization. x and y axes correspond to the 3-rd and 1-st dimensions of y_a, respectively.

Figure 9: Appearance compatibility space visualization. x and y axes correspond to the 3-rd and 1-st dimensions of y_a, respectively.

Detail of the compatibility loss

We describe more details of the compatibility loss mentioned in Section 4.2 of the main paper, which inject compatibility information for other compared methods [10, 1, 6, 7]. More specifically, similar to [9], we concatenate the generated target clothing with contextual garments to input to a discriminator. And this discriminator learns to predict {real target clothing, its contextual garments} as real, and predict both {fake target clothing, its contextual garments} and {real target clothing, incompatible (real but mismatched) contextual garments} as fake. The mismatched contextual garments are from other images according to our weakly-supervised compatibility assumption. Thus, the generated clothing is enforced to match its contextual garments. More details can be found in Section 4.2 of [9]. For our own baselines (*i.e.*, FiNet w/o 2-stage w/o comp, FiNet w/o comp, FiNet w/o two-stage), we do not have this compatibility loss and use the similar KL regularization as our full method.

More Qualitative Ablations.

We show several more visual results comparing FiNet with its baselines in Figure 10. FiNet w/o two-stage fails to handle clothing boundaries. FiNet w/o two-stage w/o comp does not generate clothing that matches the contextual garments.

Figure 10: Visual comparisons of FiNet and its ablations.

Clothing Reconstruction and Transfer

Although clothing reconstruction and transfer is not our main contribution, we show more transfer results in Figure 11 and 12 to demonstrate that our method, by reconstructing a target garment and fill it in the missing regions of an input image, can transfer the target garment naturally to the input image. Note that FiNet transfers shape and appearance by inpainting a specific clothing item, which is different from most existing approaches that generate the full person as a whole [8, 1, 3]. This potentially provides a new solution for applications like virtual try-on [3] and generating people in diverse clothes [7].

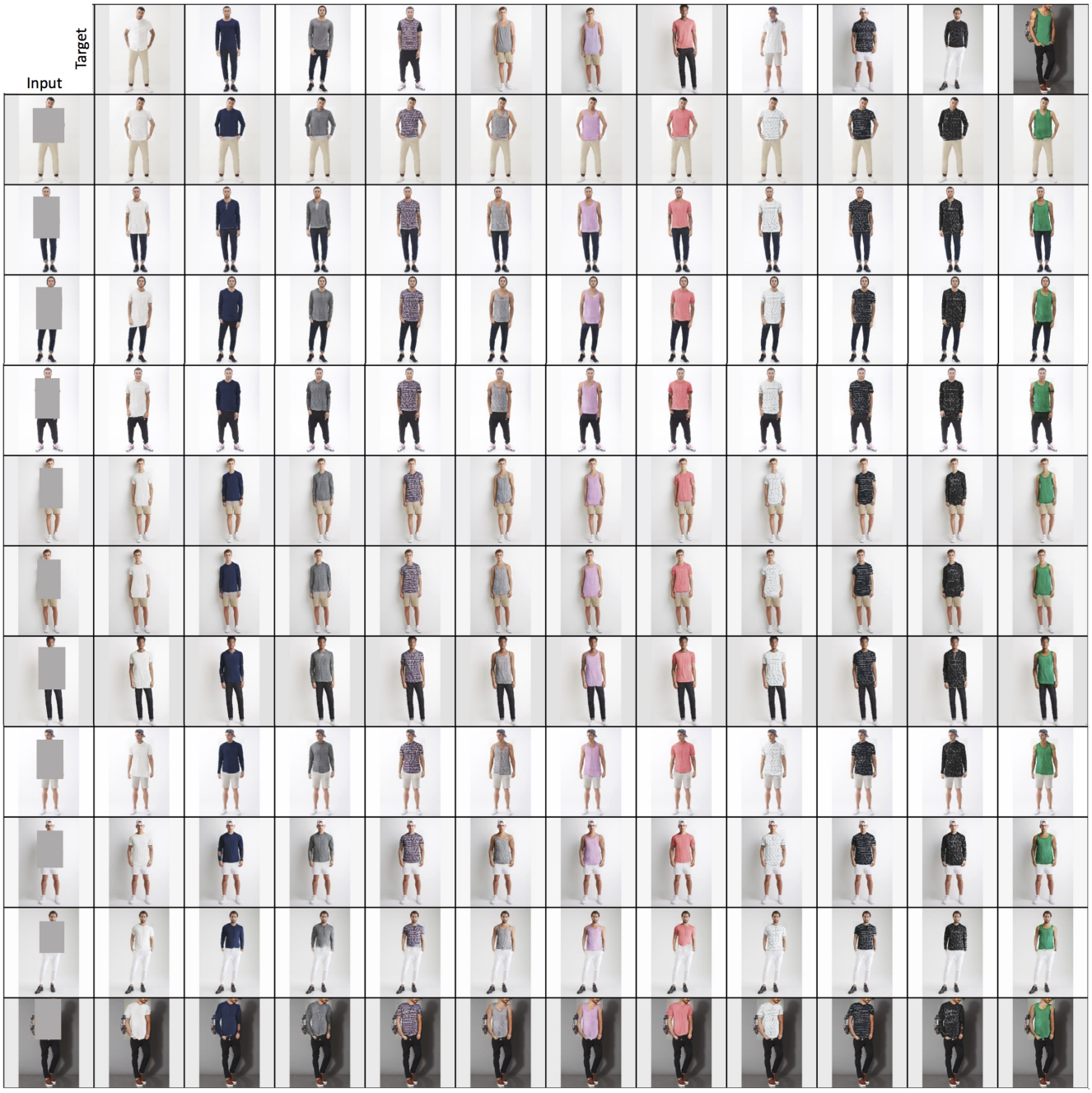

Figure 11: Clothing transfer results of tops. Each row corresponds to an input image whose top garment is transfered from different target tops. The diagonal images are reconstruction results, since the input and target images are the same. FiNet can naturally render the shape and appearance of the target garment onto other people with various poses and body shapes.

Figure 12: Clothing transfer results of bottoms. Each row corresponds to an input image whose bottom garment is transfered from different target bottoms. The diagonal images are reconstruction results, since the input and target images are the same. FiNet can naturally render the shape and appearance of the target garment onto other people with various poses and body shapes.

References

[1] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 1, 12, 13

[2] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 1

[3] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 13

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[5] Hengyuan Hu and Ruslan Salakhutdinov. Learning deep generative models with discrete latent variables. 2018. 5

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 12

[7] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *ICCV*, 2017. 1, 12, 13

[8] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 13

[9] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 12

[10] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, 2017. 1, 8, 12