This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Self-supervised Correlation Mining Network for Person Image Generation

Zijian Wang[1], Xingqun Qi[1], Kun Yuan[2], Muyi Sun[3†]
[1]School of AI/Auto, Beijing University of Posts and Telecommunications
[2]Kuaishou Technology [3]CRIPAC, Institute of Automation, Chinese Academy of Sciences
{wangzijianbupt, XingqunQi}@bupt.edu.cn, yuankun03@kuaishou.com,
muyi.sun@cripac.ia.ac.cn

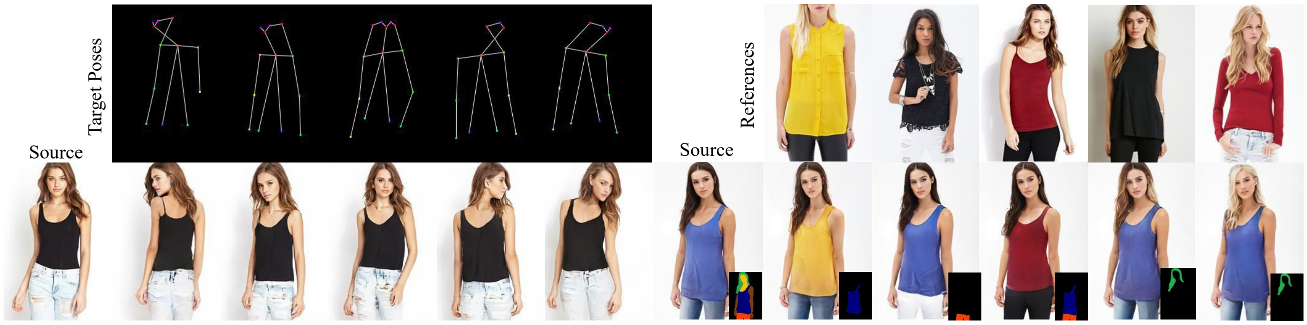

Figure 1. Exemplary samples synthesized by our self-supervised person image generation framework. Our method could generate the same person images with the target poses (left image set), or generate person images with specific attributes referring to different person images (right image set). Better viewed by zooming in the electronic version.

Abstract

Person image generation aims to perform non-rigid deformation on source images, which generally requires unaligned data pairs for training. Recently, self-supervised methods express great prospects in this task by merging the disentangled representations for self-reconstruction. However, such methods fail to exploit the spatial correlation between the disentangled features. In this paper, we propose a Self-supervised Correlation Mining Network (SCM-Net) to rearrange the source images in the feature space, in which two collaborative modules are integrated, Decomposed Style Encoder (DSE) and Correlation Mining Module (CMM). Specifically, the DSE first creates unaligned pairs at the feature level. Then, the CMM establishes the spatial correlation field for feature rearrangement. Eventually, a translation module transforms the rearranged features to realistic results. Meanwhile, for improving the fidelity of cross-scale pose transformation, we propose a graph based Body Structure Retaining Loss (BSR Loss) to preserve reasonable body structures on half body to full body generation. Extensive experiments conducted on DeepFashion dataset demonstrate the superiority of our method compared with other supervised and unsupervised approaches. Furthermore, satisfactory results on face generation show the versatility of our method in other deformation tasks.

1. Introduction

Pose guided person image generation is an unaligned image to image translation problem, which aims to change the posture of a person image given target poses as condition [18, 20, 21, 31, 36, 41]. Person image generation has shown great potential in many fields, such as film industry and multimedia creation. However, the difficulty of performing non-rigid deformation makes this task an active topic in the community of computer vision.

Due to the large spatial misalignment between the source and target images, existing approaches generally need unaligned data pairs to supervise the training process [18, 21, 31, 36, 41]. For instance, [31, 41] calculate the attention map between paired poses to guide the anomalous pose deformation. [3, 7, 25] establish the coordinate offset flow to promote the position-level source feature sampling for person feature alignment. With such attention or flow mechanism, the generative methods could be capable to perform spatial transformations when the source images and target poses are

[†] Corresponding author.

provided. However, collecting paired data requires heavy workload and limits the application scenarios of these supervised approaches. Therefore, some unsupervised methods are proposed to deal with this limitation [22, 30], which utilize cycle consistent methods or create pseudo labels to promote the training procedure. However, such methods still have limitations in generation quality intuitively.

Recently, self-supervised methods demonstrate powerful prospects to perform non-rigid spatial transformations with only source images [4, 19, 20]. They could learn disentangled representations of different image types, which are merged in the following for self-reconstruction. Early studies [4, 19] employ multi-branch network to disentangle different features and concatenate them to reconstruct the source images. Ma *et al*. [20] utilize AdaIN [11] for feature merging by transferring statistics from style features to pose features. However, these methods still encounter three challenges. First, the disentangled features are aligned in the feature space, which cannot provide enough supervision for spatial transformation in self-supervised methods. Second, these merging methods (*e.g*. concatenation or statistics transfer) are global operations, which are limited to exploiting the spatial correlation information. Third, the model lacks prior knowledge of invisible regions due to the self-supervised training process within the single pose scale, which limits the reasonable completions for invisible regions in the half body to full body transformation.

In this paper, we propose a Self-supervised Correlation Mining Network (SCM-Net) for person image generation. The entire architecture of SCM-Net can be summarized as disentanglement, fusion and translation. In the disentanglement phase, inspired by the decomposed strategy in [21], we design a Decomposed Style Encoder (DSE) to extract the semantic-aware decoupled style features, which could form “unaligned pairs ” with its counterpart pose features. Through this design, the source image itself could provide supervision for spatial feature deformation. In the fusion phase, we propose a Correlation Mining Module (CMM) to further exploit the spatial correlation between disentangled feature pairs. The CMM module computes the pairwise correlation between the corresponding positions of feature pairs to establish the dense spatial correlation field. Based on this correlation field, our model could align these disentangled features through spatially rearranging the style feature positions. In the translation phase, a translation generator with skip connections is introduced to transform the rearranged style features into realistic person images. The entire model is trained in an end-to-end manner.

For the lack of prior information on the lower body, we design a Body Structure Retaining Loss (BSR Loss) to capture the semantic relationships among different body parts. Thus, the model could make reasonable completions based on these relationships. Specifically, we employ the graph representation to model the semantic relationships of human body parts. In this body graph, each node represents the perceptual features of each semantic region and each edge measures the similarity between each node pair. We match the graphs between each input person image and the corresponding generated result to establish the graph based constraint, which incorporates the body semantic relationships into our model.

During inference, our model could introduce new target poses for human pose transfer, and perform reference based attribute editing through partial replacement of style features. Figure 1 shows some applications of our model.

The main contributions can be summarized as follows:

- We propose a Self-supervised Correlation Mining Network (SCM-Net) to achieve person image deformation without the supervision of unaligned data pairs.
- We design two main collaborative modules, the Decomposed Style Encoder (DSE) and the Correlation Mining Module (CMM), which could perform feature disentangling and merging for person image deformation.
- We propose a Body Structure Retaining Loss (BSR Loss) to acquire the prior knowledge of invisible regions through incorporating semantic relationships among body parts.
- Our method performs competitive results compared with the state-of-the-art methods and also obtains satisfactory results on face generation tasks, which demonstrates the migration capability of our model.

2. Related Work

2.1. Person Image Generation

With the dramatic development of Generative Adversarial Networks(GANs) [6], person image generation have made great progress in recent years [4, 18–22, 25, 30, 31, 36, 37, 41]. Ma *et al*. [18] first introduced the pose-guide person image generation task and proposed a two stage generator to generate target person image. Zhu *et al*. [31, 41] proposed an attention mechanism to transfer the image information from source pose to target pose. Ren *et al*. [25] predicted the flow field between source person images and target poses for generating new pose images. Men *et al*. [21] used decomposed component encoding strategy to achieve pose transfer and person attribute editing. However, all the above methods need paired data to supervise the training process, which would take heavy workload for data collection. Several unsupervised methods have been proposed for person image generation. Pumarola *et al*. [22] designed a bidirectional generator and employed cycle-consistent method to supervise the training. Song *et al*. [30] designed a novel schema to generate pseudo semantic maps for the unsupervised generation. However, these methods still need extra target poses as input and have some artifacts in generated

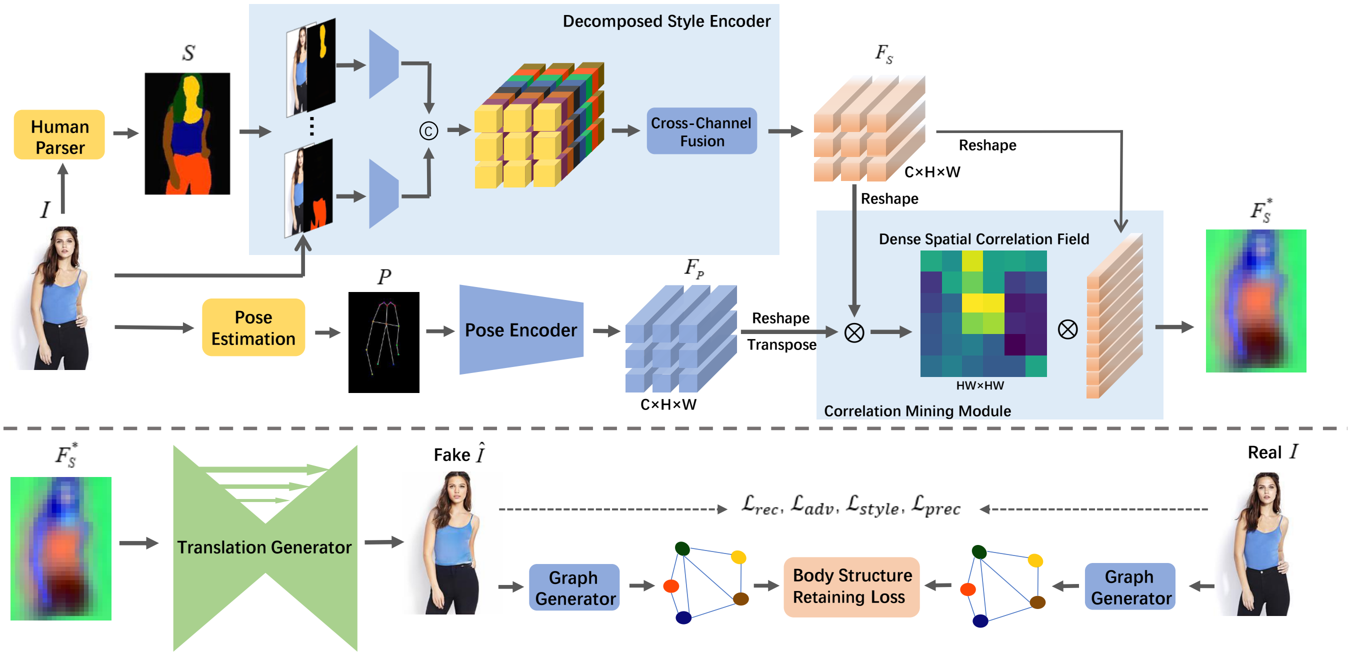

Figure 2. An overall workflow of our self-supervised person image generation framework. Given an input person image I, we first utilize the pre-trained methods to obtain its parsing map S and pose skeleton P. Then, the Decomposed Style Encoder disentangles the semantic-aware decoupled style features F_s from its pose features F_p. Next, the Correlation Mining Module establishes the correlation field $\mathcal{C}$ to guide the feature merging. Finally, the merged features F_s^* are fed into the translation generator to get the reconstruction $\hat{I}$ of the source image. The Body Structure Retaining Loss and other losses are designed to promote the training process.

images. Recently, [4, 19, 20] proposed self-driven methods to settle these problems. However, these methods have limitations when dealing with large pose deformation problems. Inspired by the above, in this paper, we propose a novel self-supervised framework with graph representation learning for person image generation.

2.2. Spatial Correlation Learning

The purpose of spatial correlation learning is to establish dense spatial correlation fields for image translation. Liao *et al*. [16] proposed a coarse-to-fine strategy to compute the spatial correlation field for image analogy and style transfer. He *et al*. [8] measured the spatial similarity between the reference and the target to perform exemplar-based colorization. Lee *et al*. [14] designed a spatially correlation related module to introduce information from reference image to sketch image for sketch colorization. Zhang *et al*. [36] proposed a spatial-aware normalization module to preserve the spatial context relationship for human pose transfer. Zhang *et al*. [37] established the spatial correlation field in a shared domain to perform cross-domain image to image translation. However, the above methods can only handle the spatial correlation between unaligned data pairs. In this paper, we establish the correlation field between the disentangled features of source images, which explores more scenarios for spatial correlation learning.

2.3. Graph Representation Learning

Graph representation learning plays a significant role in the computer vision [1, 34, 39]. Due to the powerful capabilities of relationship modeling, the graph representation learning has been applied to many tasks, such as skeleton-based action recognition [34], biometrics recognition [24] and person re-identification [28,33,35]. Yan *et al*. [35] built a person-feature based graph to model the relations among images for person search. Ren *et al*. [24] proposed a dynamic graph for occlusion biometrics recognition. Wu *et al*. [33] proposed an adaptive graph representation learning scheme to promote the interactions between relevant regional features for video person Re-ID. Hou *et al*. [10] proposed a graph matching strategy to distill structural knowledge for road marking segmentation. Qi *et al*. [23] proposed an adaptive re-weighting graph to balance the contributions of different semantic nodes in face sketch synthesis. However, the above methods employ graph representation learning to enhance the ability for feature extraction or feature matching, ignoring of the characteristics of graphs for cross-scale image complication. In this paper, we apply graph representation to model the semantic relationships for person image generation, aiming to generate more reasonable body structures.

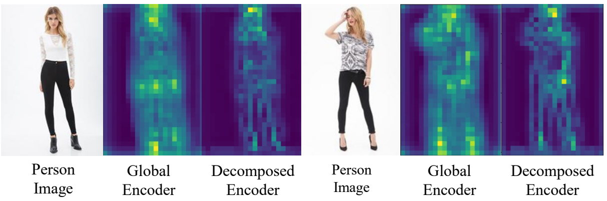

Figure 3. Feature map visualizations of global encoder and the DSE module. The structural information represented by the DSE module is significantly reduced compared with the global encoder.

3. Method

In this section, we present our proposed method in detail. To begin with, we introduce the overall workflow of our Self-supervised Correlation Mining Network (SCM-Net). Then we describe the whole network architecture in detail according to the three phases of disentangling, merging and translation. Finally, the total objective functions of our model are introduced.

3.1. Overall Workflow

Without requiring unaligned data pairs, our method receives a single source image as input. As shown in Figure 2, given a source person image I, we leverage the pre-trained human pose estimation model [2] and human parser [5] to obtain its pose skeleton P and semantic mask S. For feature disentangling, we employ the pose encoder and the DSE module to extract the pose features $F_p \in \mathbb{R}^{C\times H\times W}$ and the semantic-aware decoupled style features $F_s \in \mathbb{R}^{C\times H\times W}$, respectively. For feature merging, the CMM module is proposed to establish the dense spatial correlation field $\mathcal{C}$. Based on this correlation field, the F_s could perform spatial rearrangement to obtain the merged features F_s^*. Eventually, the translation generator G transforms the F_s^* from the feature domain to realistic images.

3.2. Disentangled Feature Encoding

For feature disentangling, there are two branches (*e.g.*, pose branch, style branch) in our framework to encode pose features and style features, respectively.

Pose Encoding. In the pose encoding branch, we employ the down-sampling convolutional neural networks (CNNs) to extract pose feature maps F_p from the pose skeleton P. Since the F_p is globally encoded, its structure is aligned with source image I inherently.

Decomposed Style Encoding. For style encoding, we design the DSE module to obtain the semantic-aware decoupled style features F_s, which could form "unaligned data pairs" with their counterpart pose features. Compared with the global encoder which encodes the person image entirely, the DSE module could embed the person image I from a complex manifold to the feature space according to different regions.

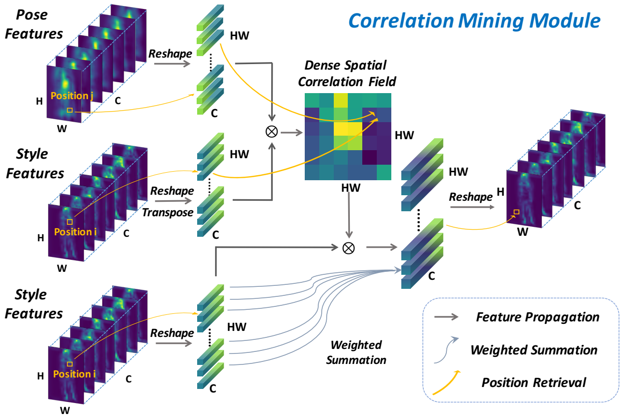

Figure 4. Details of the Correlation Mining Module in our model. Each position of outputs is the weighted average summation of the input. The weights are stored in the correlation field.

As illustrated in Figure 2, we separate the segmentation map S into 8-channel binary masks. Each channel indicates a specific body region (*e.g.*, pants, hair). Then we employ element-wise multiplication between each binary mask and the source person image I to obtain body parts. In addition, we feed each body part into an encoder whose parameters are shared for all regions to extract the partial style features $F_s^i, i \in [1, 8]$. Finally, we concatenate all F_s^i along the channel dimension to construct the semantic-aware decoupled style features F_s. Each position in style feature maps contains specific semantic information. Furthermore, for eliminating the limitation caused by the fixed concatenation order, we propose a Cross Channel Fusion (CCF) module to endow plentiful information into each position by selecting desired semantic features from different semantic regions. In structure, the CCF module has a concise design which consists of two 1×1 convolutional blocks.

To verify the effect of DSE, we visualize the feature maps extracted by the global encoder and the DSE, respectively. As shown in Figure 3, we can observe that the signal strength distribution of the global encoder represents the structural information clearly, while the distribution of the DSE is relatively flat, which indicates the structural information degradation.

3.3. Correlation based Feature Merging

In the merging phase, we propose the CMM module, which aims to establish the dense spatial correlation field $\mathcal{C}$ for feature rearrangement. To begin with, we reshape the feature F_i into $[F_i(1), F_i(2), \cdots, F_i(hw)] \in \mathbb{R}^{C\times HW}, i \in \{p, s\}$. Each vector $F_i(j) \in \mathbb{R}^C$ in F_i represents the semantic information of the j^{th} location in the feature map,

$j \in [1, hw]$.

As illustrated in Figure 4, given F_s and F_p, each vector $F_s(i)$ serves as the query to retrieve the relevant key $F_p(j)$ from the F_p. Therefore, the correlation field $\mathcal{C} \in \mathbb{R}^{HW \times HW}$ is established, whose element C_{ij} is the correlation of key-value pairs, followed by a softmax activation.

$$C_{ij} = \frac{exp(s_{ij})}{\sum_{i=1}^{hw} exp(s_{ij})} \tag{1}$$

$$s_{ij} = \frac{\bar{F}_s(i)\bar{F}_p(j)}{||\bar{F}_s(i)||\ ||\bar{F}_p(j)||} \tag{2}$$

where $\bar{F}_s(i)$ and $\bar{F}_p(j)$ represent the centralized feature, *i.e.* $\bar{F}_s(i)$ = $F_s(i)$ - mean($F_s(i)$). The correlation field $\mathcal{C}$ contains the weights which could be assigned to value vectors for feature rearrangement. Specifically, the rearranged feature $F_s^* = [F_s^*(1), F_s^*(2), \cdots, F_s^*(hw)] \in \mathbb{R}^{C \times HW}$ is obtained by calculating the weighted average summation of all positions in feature F_s.

$$F_s^*(i) = \sum_{j=1}^{hw} c_{ij} F_s(j), i \in [1, hw] \tag{3}$$

Based on the above operations, the F_s^* is structurally aligned with the input pose which could be fed into the translation generator to synthesize a realistic person image.

3.4. Aligned Feature Translation

With the rearranged features F_s^* as input, the translation generator could synthesize the target image $\hat{I}$ for self-reconstruction. To better preserve the structural information, we employ the U-Net architecture [26] as our translation generator, as its skip connection propagates the information directly from encoder to decoder.

3.5. Objective Functions

Adversarial Learning. Following the configuration of [41], we employ two discriminators, one is a pose discriminator D_p to maintain the pose consistency, and the other is a style discriminator D_s to maintain the style consistency. Both of them promote the generator G to generate realistic images. The adversarial loss L_{adv} is listed as follows:

$$\begin{aligned} \mathcal{L}_{adv} &= \mathbb{E}_{I,P}[\log(D_s(I) \cdot D_p(I, P))] \\ &+ \mathbb{E}_{I,P}[\log((1 - D_s(G(I, P))) \\ &\cdot (1 - D_p(G(I, P)))))] \end{aligned} \tag{4}$$

Self-supervised Reconstruction. The reconstruction loss L_{rec} can be formulated as the L1 distance between the source image I and generated image $\hat{I}$, which encourages the $\hat{I}$ to be similar with the I at the pixel level.

$$\mathcal{L}_{rec} = ||\hat{I} - I||_1 \tag{5}$$

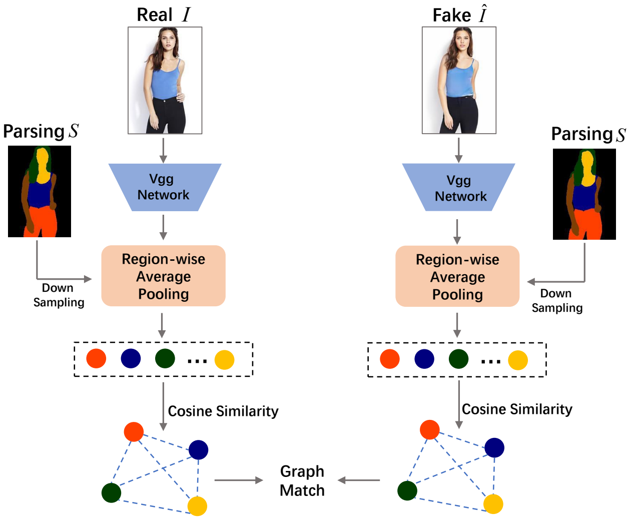

Figure 5. Details of the graph generator in our model. Nodes represent per-region styles and the edges measure the similarities between nodes.

Perceptual Consistency. The perceptual loss L_{perc} calculates the L_1 distance between the pre-trained VGG features of I and $\hat{I}$, which measures the high-level semantic differences between images [12].

$$\mathcal{L}_{perc} = ||\phi^l(\hat{I}) - \phi^l(I)||_1 \tag{6}$$

Style Consistency. The style loss L_{style} calculates the statistical errors between the pre-trained VGG features of I and $\hat{I}$, which penalizes the difference in colors and textures [12]. As shown in Formula (7), ϕ^l is the activation at the jth layer of the pre-trained VGG network, and $\mathbb{G}$ is the Gram matrix.

$$\mathcal{L}_{style} = \sum_l ||\mathbb{G}(\phi^l(\hat{I})) - \mathbb{G}(\phi^l(I))||_1 \tag{7}$$

Body Structure Retaining. The BSR Loss is proposed to endow prior knowledge of invisible regions through constraining semantic relationships among body parts. We design a graph generator to model this relationship. As illustrated in Figure 5, we employ a pre-trained VGG network and the region-wise average pooling layer [40] to obtain the body graph $\mathbb{M}$, in which the nodes represent per-region styles and the edges measure the similarities between nodes.

Due to the training process is self-supervised with the single pose in each iteration, the model cannot make a reasonable completion for the unknown regions when performing cross-scale pose transformation. Applying BSR Loss for training encourages the output person image to retain a reasonable structure, which is conductive for half body to

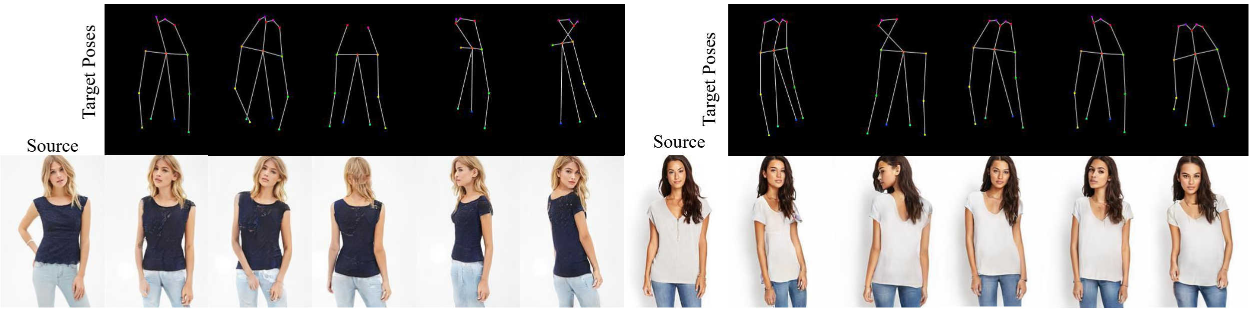

Figure 6. The results of our method in the pose guided person image generation.

full body transformation. We calculate the BSR loss between I and $\hat{I}$ as the L_{graph}.

$$\mathcal{L}_{graph} = ||\mathbb{M}(I,S) - \mathbb{M}(\hat{I},S)||_1 \tag{8}$$

The overall objective function is shown in Formula (9), where α_{adv}, α_{rec}, α_{perc}, α_{style}, α_{garph} are the weights of the corresponding loss functions.

$$\begin{aligned}\mathcal{L}_{total} &= \alpha_{adv}\mathcal{L}_{adv} + \alpha_{rec}\mathcal{L}_{rec} + \alpha_{perc}\mathcal{L}_{perc} \\ &+ \alpha_{style}\mathcal{L}_{style} + \alpha_{graph}\mathcal{L}_{graph}\end{aligned} \tag{9}$$

4. Experiments

4.1. Implementation Details

Dataset. We carry out our experiments on DeepFashion In-shop Clothes Retrieval Benchmark [17], which contains 52,712 high quality person images. We split the dataset following the same configurations of [20],

Metrics. We use the common metrics such as Structural Similarity (SSIM) [32], Inception Score (IS) [27], Learned Perceptual Image Patch Similarity (LPIPS) [38], and Fréchet Inception Distance (FID) [9] to assess the quality of generated images quantitatively. SSIM indicate the similarity between paired images in raw pixel space. Meanwhile, LPIPS, IS and FID measure the realism of the generated images at the feature level.

Network Architecture and Training Details. Both the pose encoder and style encoder employ several downsampling convolutional layers to extract features. The feature maps with resolutions 32 × 32 are applied for establishing the correlation field. Our method is implemented on PyTorch framework using 4 Nvidia TitanX GPUs. The weights for loss functions are set to $\alpha_{adv} = 5$, $\alpha_{rec} = 1$, $\alpha_{perc} = 1$, $\alpha_{style} = 150$, $\alpha_{garph} = 1$, respectively.

4.2. Pose Guided Person Image Generation

Pose guided person image generation, or pose transfer, aims to change the posture of a person image given target poses as condition. Pose transfer is an important application of person image generation. As shown in Figure 1 (left) and Figure 6 (all), given a source person image, our model could transform it to any target pose and keep the appearance details unchanged.

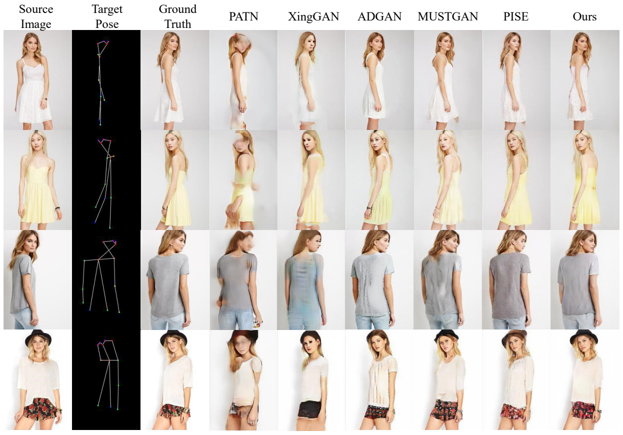

Figure 7. The comparisons with other state-of-the-art methods on pose guided person image generation. Zoom in for a better view.

Qualitative Comparison. We compare the generated images of our method with several state-of-the-art approaches, including PATN [41], XingGAN [31], ADGAN [21], MUSTGAN [20] and PISE [36]. All the results are obtained using the source code or the pre-trained model released by the authors. The results of the qualitative comparisons are shown in Figure 7. PATN and XingGAN generate blurry results since these models can not disentangle different features. The results of ADGAN and MUSTGAN have correct postures, but they fail to maintain detailed textures. This is because these models can not capture the spatial correlation well. PISE could generate desirable results. However, its results still have some unsatisfactory artifacts due to the lack of semantic relationships. Meanwhile, this model requires unaligned image pairs for training. In contrast, our model obtain competitive results only requires source images.

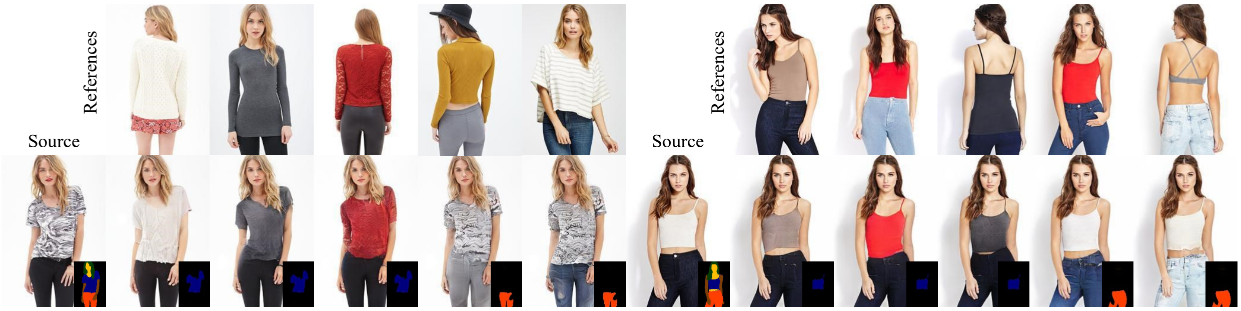

Figure 8. The results of our method in the person attribute editing.

Quantitative Comparison. As shown in Table 1, we compare our method with several state-of-the-art supervised and unsupervised methods on the DeepFashion. As we can see, our method outperforms these methods in most metrics on both supervised and unsupervised setting, which demonstrates the superiority of our method in generating high-quality person images.

Table 1. Quantitative comparisons with other supervised and unsupervised methods on DeepFashion.

Method	FID↓	SSIM↑	LPIPS↓	IS↑
Unsupervised				
VU-Net [4]	23.583	0.786	0.3211	3.087
E2E [30]	29.9	0.736	0.238	3.441
DPIG [19]	48.2	0.614	0.284	3.228
MUST [20]	15.902	0.742	-	3.692
Supervised				
Intr-Flow [15]	16.134	0.798	0.2131	3.251
Def-GAN [29]	18.547	0.770	0.2994	3.141
PATN [41]	24.071	0.770	0.2520	3.213
ADGAN [21]	18.395	0.771	0.2242	3.329
GFLA [25]	14.061	0.701	0.2219	3.635
PISE [36]	13.61	-	0.2059	-
SCM-Net	**12.18**	0.751	**0.1820**	3.632

Table 2. The evaluation results of ablation study.

Method	FID↓	SSIM↑	LPIPS↓	IS↑
w/o DSE	12.86	0.750	0.187	3.2456
w/o CCF	17.08	0.751	0.175	3.605
w/o BSR	12.61	0.755	0.178	3.441
Full	**12.18**	0.751	0.182	**3.632**

4.3. Ablation Study

We further perform the ablation study to analyze the contribution of each module and the proposed BSR Loss in our method. Firstly, we introduce the variants implemented by alternatively removing a specific component from our full model. There are four settings in this module ablation. 1). **W/o DSE.** This model removes the DSE module and directly uses a global encoder to extract the style features. 2). **W/o CCF.** This model removes the Cross Channel Fusion module from the DSE. 3). **W/o BSR.** This model removes the BSR Loss during the training procedure. 4). **Full.** This model represents our full model.

Table 2 shows the quantitative results of the ablation study. We can observe that our full model achieves the best performance on FID and IS metrics. Meanwhile, the removal of any components will degrade the performance of the model integrally. Figure 9 shows the qualitative comparisons of different ablation models. We can observe that the w/o DSE model fails to preserve the styles of source images and the w/o CCF model has limitation in preserving the detailed texture. Meanwhile, w/o BSR model can not complete the lower body well while the full model could generate reasonable results. It demonstrates that BSR Loss enhance the model's ability of capturing body structural information. Furthermore, we illustrate the comparisons on half body to full body transformation with previous self-supervised method MUST-GAN [20]. Figure 10 shows the advantages of our method when performing half body to full body transformation. We can observe that MUST-GAN [20] would generate more artifacts, while our method could complete the lower part of the body reasonably with correlation learning.

4.4. Person Attribute Editing

Our model can also achieve person attribute editing based on reference images by exchanging channel features of specific semantic areas in semantic-aware decoupled style features. As shown in Figure 1 (right) and Figure 8 (all), our method could edit the style of the upper clothes, pants and hair style respectively.

4.5. Applications on Face Generation tasks

In this section, we demonstrate the versatility of our method. Since our method could disentangle the shape and style features, it could also be applied to other image generation tasks under this self-supervised framework. Two face

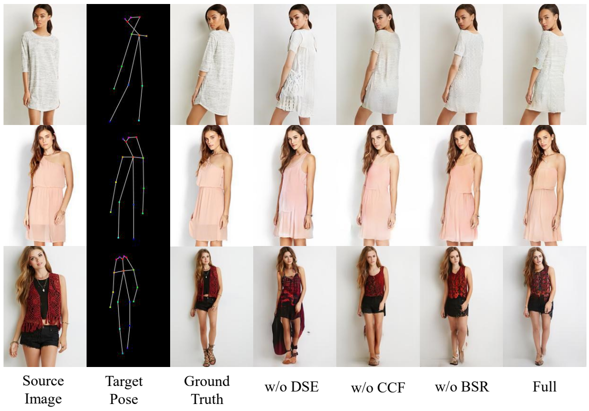

Figure 9. The qualitative comparison of ablation study.

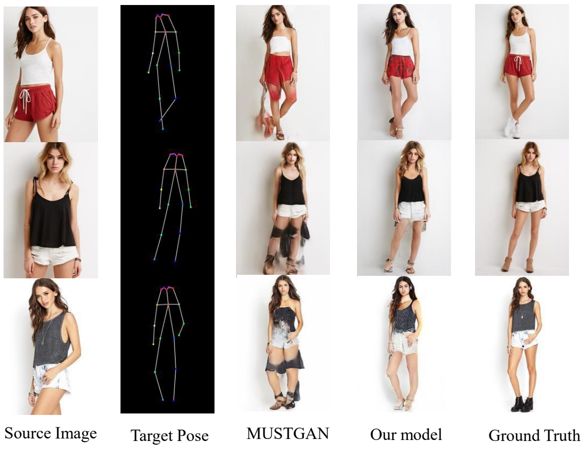

Figure 10. Results of half body to full body transformation

generation tasks are shown as follows.

Reference-Based Edge Colorization. Reference-based edge colorization aims to translate an edge map to a realistic image based on a reference image. Regarding the edge map as a pose skeleton and the reference image as a person image, our self-supervised model could achieve edge colorization. We obtain the edge map following [37] and use CelebA-HQ [13] dataset for training. The results are shown in Figure 11 (top). We can observe that the results maintain a good style consistency with the reference image, and preserve a good shape consistency with input edge maps.

Face Attribute Editing. Similar to person attribute editing, our method could also achieve face attribute editing. The results can be found in Figure 11 (bottom). We can edit specific attributes while keeping other attributes unchanged.

5. Limitation

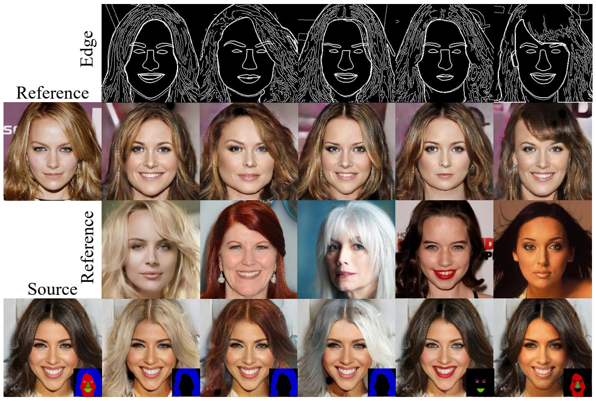

Figure 11. The results of our method in reference-based face edge colorization (top) and face attribute editing (bottom).

As shown in figure 12, our self-supervised model sometimes directly transfers certain source patterns into the final results when performing pose transfer, which is a rare situation in the supervised model. This, we hypothesis, is caused by the inherent defects of self-supervised strategy that the self-reconstruction process makes the model easy to overfit. This phenomenon might be avoided by employing spatial transformation to perform data augmentation during training in the future.

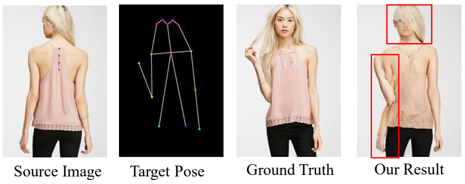

Figure 12. The illustrations of the model limitation. The hair and left arm are transferred directly from source image.

6. Conclusion

In this paper, we propose a Self-supervised Correlation Mining Network (SCM-Net) for person image generation. We propose two specially designed modules, the DSE module for feature disentanglement, and the CMM module for feature merging based on the spatial correlation. Meanwhile, the BSR Loss is proposed to promote our network to better capture the structural information, especially for half body to full body transformation. Extensive experiment results conducted on person and face datasets demonstrate the superiority of our method.

Acknowledgements: This work was supported in part by the National Key Research and Development Program of China under Grant No.2020AAA0140002, the Natural Science Foundation of China (Grant No.62076240, 62006227).

References

[1] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Deep neural networks for learning graph representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 4

[3] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *arXiv preprint arXiv:1810.11610*, 2018. 1

[4] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 2, 3, 7

[5] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 4

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[7] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 1

[8] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 3

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[10] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12486–12495, 2020. 3

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 8

[14] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5801–5810, 2020. 3

[15] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 7

[16] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 3

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[18] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017. 1, 2

[19] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018. 2, 3, 7

[20] Tianxiang Ma, Bo Peng, Wei Wang, and Jing Dong. Must-gan: Multi-level statistics transfer for self-driven person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13622–13631, 2021. 1, 2, 3, 6, 7

[21] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 1, 2, 6, 7

[22] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018. 2

[23] Xingqun Qi, Muyi Sun, Weining Wang, Xiaoxiao Dong, Qi Li, and Caifeng Shan. Face sketch synthesis via semantic-driven generative adversarial network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 3

[24] Min Ren, Yunlong Wang, Zhenan Sun, and Tieniu Tan. Dynamic graph representation for occlusion handling in biometrics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11940–11947, 2020. 3

[25] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 2, 7

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 6

[28] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 486–504, 2018. 3

[29] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 7

[30] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2357–2366, 2019. 2, 7

[31] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020. 1, 2, 6

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[33] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 29:8821–8830, 2020. 3

[34] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 3

[35] Yichao Yan, Qiang Zhang, Bingbing Ni, Wendong Zhang, Minghao Xu, and Xiaokang Yang. Learning context graph for person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2158–2167, 2019. 3

[36] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021. 1, 2, 3, 6, 7

[37] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 2, 3, 8

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[39] Shaobo Zhang, Wanqing Zhao, Ziyu Guan, Xianlin Peng, and Jinye Peng. Keypoint-graph-driven learning framework for object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1065–1073, 2021. 3

[40] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 5

[41] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 1, 2, 5, 6, 7