Unsupervised Person Image Generation with Semantic Parsing Transformation

Sijie Song[1], Wei Zhang[2], Jiaying Liu[1]*, Tao Mei[2]

[1] Institute of Computer Science and Technology, Peking University, Beijing, China
[2] JD AI Research, Beijing, China

Abstract

In this paper, we address unsupervised pose-guided person image generation, which is known challenging due to non-rigid deformation. Unlike previous methods learning a rock-hard direct mapping between human bodies, we propose a new pathway to decompose the hard mapping into two more accessible subtasks, namely, semantic parsing transformation and appearance generation. Firstly, a semantic generative network is proposed to transform between semantic parsing maps, in order to simplify the non-rigid deformation learning. Secondly, an appearance generative network learns to synthesize semantic-aware textures. Thirdly, we demonstrate that training our framework in an end-to-end manner further refines the semantic maps and final results accordingly. Our method is generalizable to other semantic-aware person image generation tasks, e.g., clothing texture transfer and controlled image manipulation. Experimental results demonstrate the superiority of our method on DeepFashion and Market-1501 datasets, especially in keeping the clothing attributes and better body shapes.

1. Introduction

Pose-guided image generation has attracted great attentions recently, which is to change the pose of the person image to a target pose, while keeping the appearance details. This topic is of great importance in fashion and art domains for a wide range of applications from image / video editing, person re-identification to movie production.

With the development of deep learning and generative model [8], many researches have been devoted to pose-guided image generation [19, 21, 5, 27, 26, 1, 20]. Initially, this problem is explored under the fully supervised setting [19, 27, 26, 1]. Though promising results have been presented, their training data has to be composed with paired images (*i.e.*, same person in the same clothing but in different poses). To tackle this data limitation and enable more

---

*Corresponding author. This work was done at JD AI Research. Our project is available at https://github.com/SijieSong/person_generation_spt.git.

Figure 1: Visual results of different methods on DeepFashion [18]. Compared with PG^2 [19], Def-GAN [27], and UPIS [21], our method successfully keeps the clothing attributes (*e.g.*, textures) and generates better body shapes (*e.g.*, arms).

flexible generation, more recent efforts have been devoted to learning the mapping with unpaired data [21, 5, 20]. However without "paired" supervision, results in [21] are far from satisfactory due to the lack of supervision. Disentangling image into multiple factors (*e.g.*, background / foreground, shape / appearance) is explored in [20, 5]. But ignoring the non-rigid human-body deformation and clothing shapes leads to compromised generation quality.

Formally, the key challenges of this unsupervised task are in three folds. First, due to the non-rigid nature of human body, transforming the spatially misaligned body-parts is difficult for current convolution-based networks. Second, clothing attributes, *e.g.*, sleeve lengths and textures, are generally difficult to preserve during generation. However, these clothing attributes are crucial for human visual perception. Third, the lack of paired training data gives little clue in establishing effective training objectives.

To address these aforementioned challenges, we propose to seek a new pathway for unsupervised person image generation. Specifically, instead of directly transforming the person image, we propose to transform the semantic parsing between poses. On one hand, translating between *person*

image and *semantic parsing* (in both directions) has been extensively studied, where sophisticated models are available. On the other hand, semantic parsing transformation is a much easier problem to handle spatial deformation, since the network does not care about the appearance and textures.

As illustrated in Fig. 2, our model for unsupervised person image generation consists of two modules: semantic parsing transformation and appearance generation. In semantic parsing transformation, a semantic generative network is employed to transform the input semantic parsing to the target parsing, according to the target pose. Then an appearance generative network is designed to synthesize textures on the transformed parsing. Without paired supervision, we create pseudo labels for semantic parsing transformation and introduce cycle consistency for training. Besides, a semantic-aware style loss is developed to help the appearance generative network learn the essential mapping between corresponding semantic areas, where clothing attributes can be well-preserved by rich semantic parsing. Furthermore, we demonstrate that the two modules can be trained in an end-to-end manner for finer semantic parsing as well as the final results.

In addition, the mapping between corresponding semantic areas inspires us to apply our appearance generative network on applications of semantic-guided image generation. Conditioning on the semantic map, we are able to achieve clothing texture transfer of two person images. In the meanwhile, we are able to control the image generation by manually modifying the semantic map.

The main contributions can be summarized as follows:

- We propose to address the unsupervised person image generation problem. Consequently, the problem is decomposed into semantic parsing transformation (H_S) and appearance generation (H_A).

- We design a delicate training schema to carefully optimize H_S and H_A in an end-to-end manner, which generates better semantic maps and further improves the pose-guided image generation results.

- Our model is superior in rendering better body shape and keeping clothing attributes. Also it is generalizable to other conditional image generation tasks, *e.g.*, clothing texture transfer and controlled image manipulation.

2. Related Work

2.1. Image Generation

With the advances of generative adversarial networks (GANs) [8], image generation has received a lot of attentions and been applied on many areas [15, 29, 4, 31].

There are mainly two branches in this research field. One lies in supervised methods and another lies in unsupervised methods. Under the supervised setting, pix2pix [11] built a conditional GAN for image to image translation, which is essentially a domain transfer problem. Recently, more efforts [15, 29] have been devoted to generating really high-resolution photo-realistic images by progressively generating multi-scale images. For the unsupervised setting, reconstruction consistency is employed to learn cross-domain mapping [34, 32, 16]. However, these unsupervised methods are developed and applied mostly for appearance generation of the spatially aligned tasks. With unpaired training data, our work is more intractable to learn the mapping to handle spatial non-rigid deformation and appearance generation simultaneously.

2.2. Pose-Guided Person Image Generation

The early attempt on pose-guided image generation was achieved by a two-stage network PG^2 [19], in which the output under the target pose is coarsely generated in the first stage, and then refined in the second stage. To better model shape and appearance, Siarohin *et al.* [27] utilized deformable skips to transform high-level features of each body part. Similarly, the work in [1] employs body part segmentation masks to guide the image generation. However, [19, 27, 1] are trained with paired data. To relieve the limitation, Pumarola *et al.* [21] proposed a fully unsupervised GAN, borrowing the ideas from [34, 22]. On the other hand, the works in [5, 20] solved the unsupervised problem by sampling from feature spaces according to the data distribution. These sample based methods are less faithful to the appearance of reference images, since they generate results from highly compressed features. Instead, we use semantic information to help preserve body shape and texture synthesis between corresponding semantic areas.

2.3. Semantic Parsing for Image Generation

The idea of inferring scene layout (semantic map) has been explored in [10, 14] for text-to-image translation. Both of the works illustrate that by conditioning on estimated layout, more semantically meaningful images can be generated. The scene layout is predicted from texts [10] or scene graphs [14] with the supervision from groundtruth. In contrast, our model learns the prediction for semantic map in an unsupervised manner. We also show that the semantic map prediction can be further refined by end-to-end training.

3. The Proposed Method

Given a target pose $\mathbf{p}_t$ and a reference image I_{p_s} under pose $\mathbf{p}_s$, our goal is to generate an output image $\tilde{I}_{p_t}$, which follows the clothing appearance of I_{p_s} but under the pose $\mathbf{p}_t$. This generation can be formulated as: $< I_{p_s}, \mathbf{p}_t > \to \tilde{I}_{p_t}$.
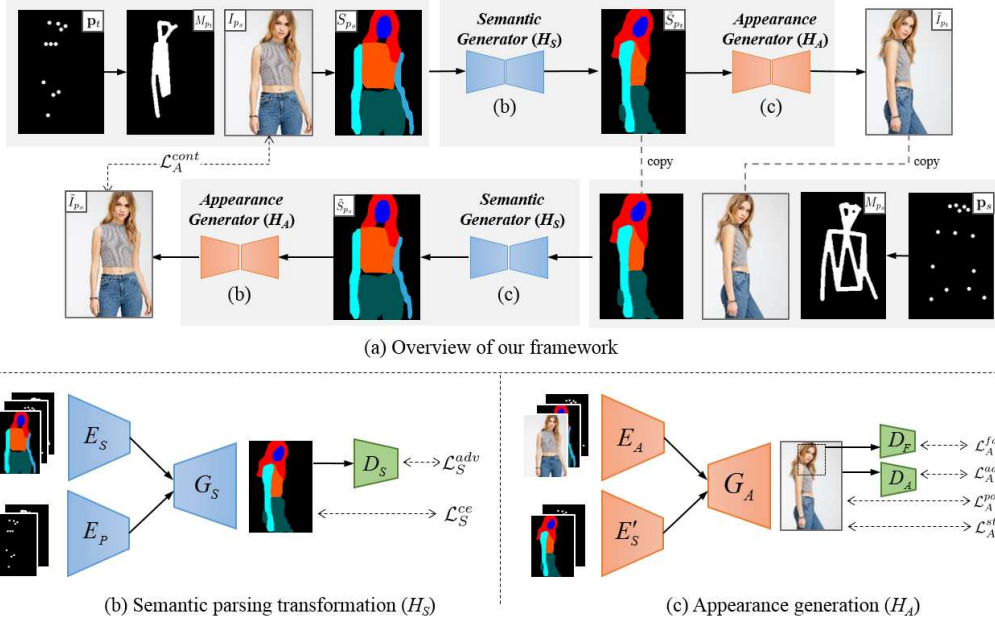
Figure 2: Our framework for unsupervised person image generation.

During the training process, we are under an unsupervised setting: the training set is composed with $\{I_{p_s}^i, \mathbf{p}_s^i, \mathbf{p}_t^i\}_{i=1}^N$, where the corresponding ground-truth image $I_{p_t}^i$ is not available. For this challenging unpaired person image generation problem, our key idea is to introduce human semantic parsing to decompose it into two modules: *semantic parsing transformation* and *appearance generation*. Our overall framework can be viewed in Fig. 2(a). Semantic parsing transformation module aims to first generate a semantic map under the target pose, which provides crucial prior for the human body shape and clothing attributes. Guided by the predicted semantic map and the reference image, appearance generation module then synthesizes textures for the final output image.

In the following, we first introduce person representation, which is the input of our framework. We then describe each module in details from the perspective of independent training. Finally, we illustrate the joint learning of the two modules in an end-to-end manner.

3.1. Person Representation

Besides the reference image $I_{p_s} \in \mathbb{R}^{3 \times H \times W}$, the source pose $\mathbf{p}_s$, and the target pose $\mathbf{p}_t$, our model also involves a semantic map S_{p_s} extracted from I_{p_s}, pose masks M_{p_s} for $\mathbf{p}_s$ and M_{p_t} for $\mathbf{p}_t$. In our work, we represent poses as probability heat maps, *i.e.*, $\mathbf{p}_s, \mathbf{p}_t \in \mathbb{R}^{k \times H \times W} (k = 18)$. The semantic map S_{p_s} is extracted with an off-the-shelf human parser [7]. We represent S_{p_s} using a pixel-level one-hot encoding, *i.e.*, $S_{p_s} \in \{0, 1\}^{L \times H \times W}$, where L indicates the total number of semantic labels. For the pose masks M_{p_s} and M_{p_t}, we adopt the same definition in [19], which provide

prior on pose joint connection in the generation process.

3.2. Semantic Parsing Transformation (H_S)

In this module, we aim to predict the semantic map $\tilde{S}_{p_t} \in [0, 1]^{L \times H \times W}$ under the target pose $\mathbf{p}_t$, according to the reference semantic map S_{p_s}. It is achieved by the semantic generative network, which is based on U-Net [23]. As shown in Fig. 2(b), our semantic generative network consists of a semantic map encoder E_S, a pose encoder E_P and a semantic map generator G_S. E_S takes S_{p_s}, $\mathbf{p}_s$ and M_{p_s} as input to extract conditional sematic information, while E_P takes $\mathbf{p}_t$ and M_{p_t} as input to encode the target pose. G_S then predicts $\tilde{S}_{p_t}$ based on the encoded features. As [35], *softmax* activation function is employed at the end of G_S to generate the semantic label for each pixel. Formally, the predicted semantic map $\tilde{S}_{p_t}$ conditioned on S_{p_s} and $\mathbf{p}_t$ is formulated as $\tilde{S}_{p_t} = G_S(E_S(S_{p_s}, \mathbf{p}_s, M_{p_s}), E_P(\mathbf{p}_t, M_{p_t}))$. The introduction of M_{p_s} and M_{p_t} as input is to help generate continuous semantic maps, especially for bending arms.

Pseudo label generation. The semantic generative network is trained to model the spatial semantic deformation under different poses. Since semantic maps do not associate with clothing textures, people in different clothing appearance may share similar semantic maps. Thus, we can search similar semantic map pairs in the training set to facilitate the training process. For a given S_{p_s}, we search a semantic map $S_{p_t^*}$ which is under different poses but shares the same clothing type as S_{p_s}. Then we use $\mathbf{p}_t^*$ as the target pose for S_{p_s}, and regard $S_{p_t^*}$ as the pseudo ground truth. We define a simple yet effective metric for the search problem. The

human body is decomposed into ten rigid body subparts as in [27], which can be represented with a set of binary masks $\{B^j\}_{j=1}^{10}(B^j \in \mathbb{R}^{H \times W})$. $S_{p_t^*}$ is searched by solving

$$S_{p_t^*} = \arg\min_{S_p} \sum_{j=1}^{10} ||B_p^j \otimes S_p - f_j(B_{p_s}^j \otimes S_{p_s})||_2^2, \quad (1)$$

where $f_j(\cdot)$ is an affine transformation to align the two body parts according to four corners of corresponding binary masks, $\otimes$ denotes the element-wise multiplication. Note that pairs sharing very similar poses are excluded.

Cross entropy loss. The semantic generative networks can be trained under supervision with paired data $\{S_{p_s}, \mathbf{p}_s, S_{p_t^*}, \mathbf{p}_t^*\}$. We use the cross-entropy loss $\mathcal{L}_S^{ce}$ to constrain pixel-level accuracy of semantic parsing transformation, and we give the human body more weight than the background with the pose mask $M_{p_t^*}$ as

$$\mathcal{L}_S^{ce} = -||S_{p_t^*} \otimes log(\tilde{S}_{p_t^*}) \otimes (1 + M_{p_t^*})||_1. \quad (2)$$

Adversarial loss. We also employ an adversarial loss $\mathcal{L}_S^{adv}$ with a discriminator D_S to help G_S generate semantic maps of *visual style* similar to the realistic ones.

$$\mathcal{L}_S^{adv} = \mathcal{L}^{adv}(H_S, D_S, S_{p_t^*}, \tilde{S}_{p_t^*}), \quad (3)$$

where $H_S = G_S \circ (E_S, E_P)$, $\mathcal{L}^{adv}(G, D, X, Y) = \mathbb{E}_X[\log D(X))] + \mathbb{E}_Y[\log(1 - D(Y)]$ and Y is associated with G.

The overall losses for our semantic generative network are as follows,

$$\mathcal{L}_S^{total} = \mathcal{L}_S^{adv} + \lambda^{ce}\mathcal{L}_S^{ce}. \quad (4)$$

3.3. Appearance Generation (H_A)

In this module, we utilize the appearance generative network to synthesize textures for the output image $\tilde{I}_{p_t} \in \mathbb{R}^{3 \times H \times W}$, guided by the reference image S_{p_s} and predicted semantic map $\tilde{S}_{p_t}$ from semantic parsing transformation module. The architecture of appearance generative network consists of an appearance encoder E_A to extract the appearance of reference image I_{p_s}, a semantic map encoder E_S' to encode the predicted semantic map $\tilde{S}_{p_t}$, and an appearance generator G_A. The architecture of appearance generative network is similar to the semantic generative network, except that we employ deformable skips in [27] to better model spatial deformations. The output image is obtained by $\tilde{I}_{p_t} = G_A \left(E_A(I_{p_s}, S_{p_s}, \mathbf{p}_s), E_S'(\tilde{S}_{p_t}, \mathbf{p}_t)\right)$, as in Fig. 2(c).

Without the supervision of ground truth I_{p_t}, we train the appearance generative network using the cycle consistency as [34, 21], in which G_A should be able to map back I_{p_s} with the generated $\tilde{I}_{p_t}$ and $\mathbf{p}_s$. We denote the mapped-back image as $\tilde{I}_{p_s}$, and the predicted segmentation map as $\tilde{S}_{p_s}$ in the process of mapping back.

Adversarial loss. Discriminator D_A is first introduced to distinguish between the realistic image and generated image, which leads to adversarial loss $\mathcal{L}_A^{adv}$

$$\mathcal{L}_A^{adv} = \mathcal{L}^{adv}(H_A, D_A, I_{p_s}, \tilde{I}_{p_t}) + \mathcal{L}^{adv}(H_A, D_A, I_{p_s}, \tilde{I}_{p_s}), \quad (5)$$

where $H_A = G_A \circ (E_A, E_S')$.

Pose loss. As in [21], we use pose loss $\mathcal{L}_A^{pose}$ with a pose detector $\mathcal{P}$ to generate images faithful to the target pose

$$\mathcal{L}_A^{pose} = ||\mathcal{P}(\tilde{I}_{p_t}) - \mathbf{p}_t||_2^2 + ||\mathcal{P}(\tilde{I}_{p_s}) - \mathbf{p}_s||_2^2. \quad (6)$$

Content loss. Content loss $\mathcal{L}_A^{cont}$ is also employed to ensure the cycle consistency

$$\mathcal{L}_A^{cont} = ||\Lambda(\tilde{I}_{p_s}) - \Lambda(I_{p_s})||_2^2, \quad (7)$$

where $\Lambda(I)$ is the feature map of image I of *conv2_1* layer in VGG16 model [28] pretrained on ImageNet.

Style loss. It is challenging to correctly transfer the color and textures from I_{p_s} to $\tilde{I}_{p_t}$ without any constraints, since they are spatially misaligned. [21] tried to tackle this issue with patch-style loss, which enforces that texture around corresponding pose joints in I_{p_s} and $\tilde{I}_{p_t}$ are similar. We argue that patch-style loss is not powerful enough in two-folds: (1) textures around joints would change with different poses, (2) textures of main body parts are ignored. Another alternative is to utilize body part masks. However, they can not provide texture contour. Thanks to the guidance provided by semantic maps, we are able to well retain the style with a semantic-aware style loss to address the above issues. By enforcing the style consistency among I_{p_s}, $\tilde{I}_{p_t}$ and $\tilde{I}_{p_s}$, our semantic-aware style loss is defined as

$$\mathcal{L}_A^{sty} = \mathcal{L}^{sty}(I_{p_s}, \tilde{I}_{p_t}, S_{p_s}, \tilde{S}_{p_t}) + \mathcal{L}^{sty}(\tilde{I}_{p_t}, \tilde{I}_{p_s}, \tilde{S}_{p_t}, \tilde{S}_{p_s}), \quad (8)$$

where

$$\mathcal{L}^{sty}(I_1, I_2, S_1, S_2)$$
$$= \sum_{l=1}^{L} ||\mathcal{G}(\Lambda(I_1) \otimes \Psi_l(S_1)) - \mathcal{G}(\Lambda(I_2) \otimes \Psi_l(S_2)))||_2^2.$$

And $\mathcal{G}(\cdot)$ denotes the function for Gram matrix [6], $\Psi_l(S)$ denotes the downsampled binary map from S, indicating pixels that belong to the l-th semantic label.

Face loss. Besides, we add a discriminator D_F for generating more natural faces,

$$\mathcal{L}_A^{face} = \mathcal{L}^{adv}(H_A, D_F, \mathcal{F}(I_{p_s}), \mathcal{F}(\tilde{I}_{p_t})) + \mathcal{L}^{adv}(H_A, D_F, \mathcal{F}(I_{p_s}), \mathcal{F}(\tilde{I}_{p_s})), \quad (9)$$

where $\mathcal{F}(I)$ represents the face extraction guided by pose joints on faces, which is achieved by a non-parametric spatial transform network [12] in our experiments.

The overall losses for our appearance generative network are as follows,

$$\mathcal{L}_A^{total} = \mathcal{L}_A^{adv} + \lambda^{pose}\mathcal{L}_A^{pose} + \lambda^{cont}\mathcal{L}_A^{cont} + \lambda^{sty}\mathcal{L}_A^{sty} + \mathcal{L}_A^{face}. \quad (10)$$

3.4. End-to-End Training

Since the shape and contour of our final output is guided by the semantic map, the visual results of appearance generation rely heavily on the quality of predicted semantic map from semantic parsing transformation. However, if they are independently trained, two reasons might lead to instability for H_S and H_A.

- Searching error: the searched semantic maps are not very accurate, as in Fig. 3(a).
- Parsing error: the semantic maps obtained from human parser are not accurate, since we do not have labels to finetune the human parser, as in Fig. 3(b).

Our training scheme is shown in Algorithm 1.

Algorithm 1 End-to-end training for our network.

Input: $\{S_{p_s}^i, \mathbf{p}_s^i, S_{p_t^*}^i, (\mathbf{p}_t^*)^i\}_{i=1}^{N^*}, \{I_{p_s}^i, \mathbf{p}_s^i, \mathbf{p}_t^i\}_{i=1}^{N}$.

1: Initialize the network parameters.
    //Pre-train H_S
2: With $\{S_{p_s}^i, \mathbf{p}_s^i, S_{p_t^*}^i, (\mathbf{p}_t^*)^i\}_{i=1}^{N^*}$, train $\{H_S, D_S\}$ to optimize $\mathcal{L}_S^{total}$.
    //Train H_A
3: With $\{I_{p_s}^i, \mathbf{p}_s^i, \tilde{S}_{p_t}^i, \mathbf{p}_t^i\}_{i=1}^{N}$ and $\{H_S, D_S\}$ fixed, train $\{H_A, D_A, D_{face}\}$ to optimize $\mathcal{L}_A^{total}$.
    //Joint optimization
4: Train $\{H_S, D_S, H_A, D_A, D_{face}\}$ jointly with $\mathcal{L}_A^{total}$, using $\{I_{p_s}^i, \mathbf{p}_s^i, \tilde{S}_{p_t}^i, \mathbf{p}_t^i\}_{i=1}^{N}$.

Output: H_S, H_A.

(a) Searching error   (b) Parsing error

Figure 3: Errors exist in the searched semantic map pairs, which might cause the inaccuracy of semantic parsing transformation.

4. Experiments

In this section, we evaluate our proposed framework with both qualitative and quantitative results.

4.1. Datasets and Settings

DeepFashion [18]. We experiment with the *In-shop Clothes Retrieval Benchmark* of the DeepFashion dataset. It contains a large number of clothing images with various appearance and poses, the resolution of which is 256×256. Since our method does not require paired data, we randomly select $37,258$ images for training and $12,000$ images for testing.

Market-1501 [33]. This dataset contains 32,668 images from different viewpoints. The images are in the resolution
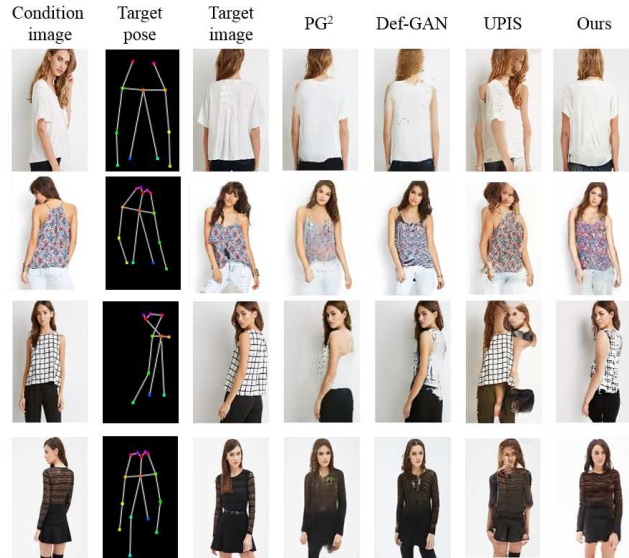

Figure 4: Example results by different methods (PG^2 [19], Def-GAN [27] and UPIS [21]) on DeepFashion. Our model better keeps clothing attributes (*e.g.*, textures, clothing types).

of 128×64. We adopt the same protocol for data split as in [33]. And we select 12,000 pairs for testing as in [27].

Implementation details. For the person representation, the 2D poses are extracted using OpenPose [2], and the condition semantic maps are extracted with the state-of-the-art human parser [7]. We integrate the semantic labels originally defined in [7] and set $L = 10$ (*i.e.*, background, face, hair, upper clothes, pants, skirt, left/right arm, left/right leg). For DeepFashion dataset, the joint learning to refine semantic map prediction is performed on the resolution of 128×128. Then we upsample the predicted semantic maps to train images in 256×256 with progressive training strategies [15]. For Market-1501, we directly train and test on 128×64. Besides, since the images in Market-1501 are in low resolution and the face regions are blurry. $\mathcal{L}_A^{face}$ is not adopted on Market-1501 for efficiency. For the hyper-parameters, we set λ^{pose}, λ^{cont} as 700, 0.03 for DeepFashion and 1, 0.003 for Market-1501. λ^{sty} is 1 for all experiments. We adopt ADAM optimizer [17] to train our network with a learning rate 0.0002 ($\beta_1 = 0.5$ and $\beta_2 = 0.999$). The batch sizes for DeepFashion and Market-1501 are set to 4 and 16, respectively. For more detailed network architecture and training scheme on each dataset, please refer to our supplementary.

4.2. Comparison with State-of-the-Arts

Qualitative Comparison. In Fig. 1, Fig. 4 and Fig. 5, we present the qualitative comparison with three state-of-the-art methods: PG^2 [19], Def-GAN [27] and UPIS [21][1].

---
[1]The results for PG^2 and Def-GAN are obtained by public models released by their authors, and UPIS are based on our implementation.

Figure 5: Example results by different methods (PG^2 [19], Def-GAN [27] and UPIS [21]) on Market-1501. Our model generates better body shapes.

PG^2 [19] and Def-GAN [27] are supervised methods that require paired training data. UPIS [21] is under the unsupervised setting, which essentially employs CycleGAN [34]. Our model generates more realistic images with higher visual quality and less artifacts. As shown in Fig. 4, our method is especially superior in keeping the clothing attributes, including textures and clothing type (the last row). Similarly in Fig. 5, our method better shapes the legs and arms. More generated results can be found in our supplementary.

Quantitative Results. In Table 1, we use the Inception Score (IS) [24] and Structural SIMilarity (SSIM) [30] for quantitative evaluation. For Market-1501 dataset, to alleviate the influence of background, mask-IS and mask-SSIM are also employed as in [19], which exclude the background area when computing IS and SSIM. For a fair comparison, we mark the training data requirements for each method. Overall, our proposed model achieves the best IS value on both datasets, even compared with supervised methods, which is in agreement with more realistic details and better body shape in our results. Our SSIM score is slightly lower than other methods, which can be explained by the fact that blurry images always achieve higher SSIM but being less photo-realistic, as observed in [20, 19, 13, 25]. Limited by space, please refer to our supplementary for user study.

4.3. Ablation Study

We design the following experiments with different configurations to first evaluate the introduction of semantic information for unpaired person image generation:

• Baseline: our baseline model without the introduction of semantic parsing, the architecture of which is the same as appearance generative network, but without semantic map as input. To keep the style on the output image, we use mask-style loss, which replaces semantic maps with body part masks in Eq. (8).

• TS-Pred: The semantic and appearance generative networks are trained independently in a two-stage manner. And we feed the predicted semantic maps into appearance generative network to get the output.

• TS-GT: The networks are trained in two-stage. We regard semantic maps extracted from target images as ground truth, and feed them into appearance generative network to get the output.

• E2E (Ours): jointly training the networks in an end-to-end manner.

Fig. 6 presents the intermediate semantic maps and the corresponding generated images. Table 1 further shows the quantitative comparisons. Without the guidance of semantic maps, the network is difficult to handle the shape and appearance at the same time. The introduction of semantic parsing transformation consistently outperforms our baseline. When trained in two-stage, the errors in the predicted semantic maps lead to direct image quality degradation. With end-to-end training, our model is able to refine the semantic map prediction. For example, the haircut and sleeves length in Fig. 6(a) are well preserved. For DeepFashion, the end-to-end training strategy leads to comparable results with that using GT semantic maps. For Market-1501, our model (E2E) achieves even higher IS and SSIM values than TS-GT. This is mainly because the human parser [7] does not work very well on low-resolution images and many errors exists in the parsing results, as the first row in Fig. 6(b).

We then analyze the loss functions in the appearance generation as shown in Fig. 7. We mainly explore the proposed style loss and face adversarial loss, since other losses are indispensable to ensure the cycle consistency. We adopt TS-GT model here to avoid the influence of semantic map prediction. In (a) and (b), we replace the semantic-aware style loss $\mathcal{L}_A^{sty}$ with mask-style loss and patch-style loss, respectively. Without semantic guidance, both of them lead to dizzy contour. Besides, the adversarial loss for faces effectively helps generate natural faces and improve the visual quality of output images.

4.4. Applications

Since the appearance generative network essentially learns the texture generation guided by semantic map, it can also be applied on other conditional image generation tasks.

Table 1: Quantitative results on DeepFashion and Market-1501 datasets (*Based on implementation).

Models	Paired data	DeepFashion		Market-1501			
		IS	SSIM	IS	SSIM	mask-IS	mask-SSIM
PG^2 [19]	Y	3.090	0.762	3.460	0.253	3.435	0.792
Def-GAN [27]	Y	3.439	0.756	3.185	0.290	3.502	**0.805**
V-Unet [5]	N	3.087	**0.786**	3.214	**0.353**	–	–
BodyROI7 [20]	N	3.228	0.614	3.483	0.099	3.491	0.614
UPIS [21]	N	2.971	0.747	3.431*	0.151*	3.485*	0.742*
Baseline	N	3.140	0.698	2.776	0.157	2.814	0.714
TS-Pred	N	3.201	0.724	3.462	0.180	3.546	0.740
TS-GT	N	3.350	0.740	3.472	0.200	3.675	0.749
E2E(Ours)	N	**3.441**	0.736	**3.499**	0.203	**3.680**	0.758
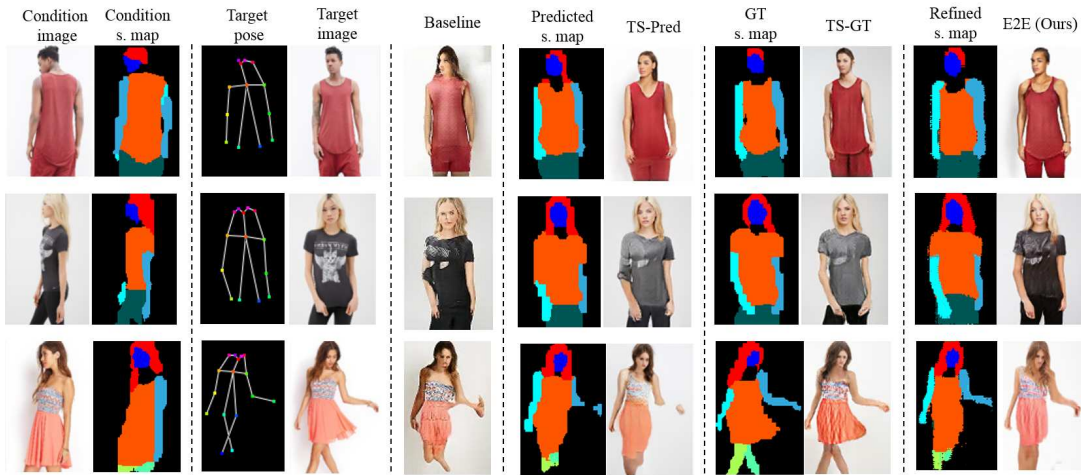
(a) Results on DeepFashion with different configurations. (Note E2E refines the haircut in the 1st row, sleeve length in the 2nd, arms in the 3rd row, compared with TS-Pred.)
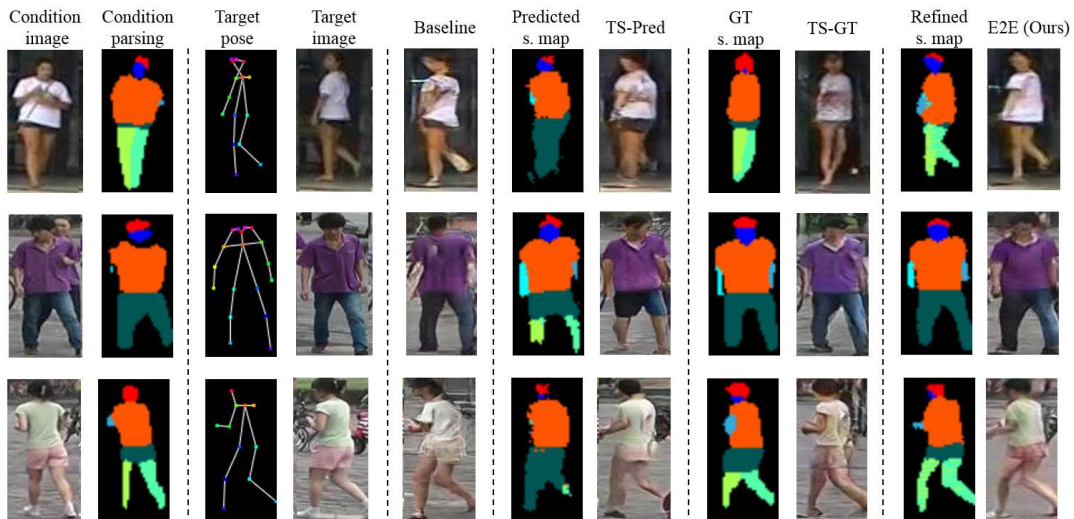
(b) Results on Market-1501 with different configurations. (Note E2E refines the body shape in the 1st and 3rd rows, pants length in the 2nd row, compared with TS-Pred.)

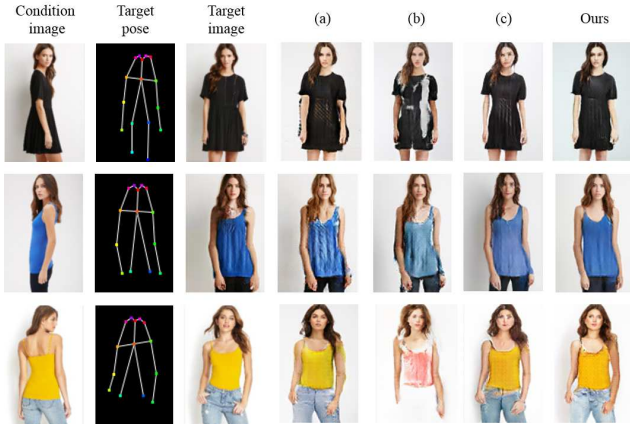Figure 6: Ablation studies on semantic parsing transformation.

Figure 7: Analysis for the loss function in appearance generation. (a) Replace $\mathcal{L}_A^{sty}$ with mask-style loss. (b) Replace $\mathcal{L}_A^{sty}$ with patch-style loss. (c) Without $\mathcal{L}_A^{face}$. Results of TS-GT with our full loss are in the right.

Figure 8: Application for clothing texture transfer. Left: condition and target images. Middle: transfer from A to B. Right: transfer from B to A. We compare our methods with image analogy [9] and neural doodle [3].

Here we show two interesting applications to demonstrate the versatility of our model.

Clothing Texture Transfer. Given the condition and target images and their semantic parsing results, our appearance generative network is able to achieve clothing texture transfer. The bidirectional transfer results can be viewed in Fig. 8. Compared with image analogy [9] and neural doodle [3], not only textures are well preserved and transferred accordingly, but also photo-realistic faces are generated automatically.

Controlled Image Manipulation. By modifying the semantic maps, we generate images in the desired layout. In Fig. 9, we edit the sleeve lengths (top), and change the dress to pants for the girl (bottom). We also compare with image analogy [9] and neural doodle [3].

4.5. Discussions for Failure Cases

Though our model generates appealing results, we show the examples of failure cases in Fig. 10. The example in the first row is mainly caused by the error in condition semantic map extracted by the human parser. The semantic genera-

Figure 9: Application for controlled image manipulation. By manually modifying the semantic maps, we can control the image generation in the desired layout.
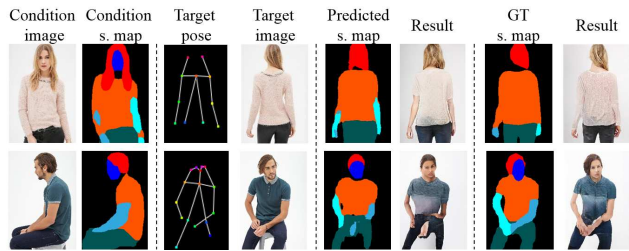

Figure 10: The failure cases in our model.

tive network is not able to predict the correct semantic map where the arms should be parsed as sleeves. The transformation in the second example is very complicated due to the rare pose, and the generated semantic map is less satisfactory, which leads to unnatural generated images. However, with groundtruth semantic maps, our model still achieves pleasant results. Thus, such failure cases can be probably solved with user interaction.

5. Conclusion

In this paper, we propose a framework for unsupervised person image generation. To deal with the complexity of learning a direct mapping under different poses, we decompose the hard task into semantic parsing transformation and appearance generation. We first explicitly predict the semantic map of the desired pose with semantic generative network. Then the appearance generative network synthesizes semantic-aware textures. It is found that end-to-end training the model enables a better semantic map prediction and further final results. We also showed that our model can be applied on clothing texture transfer and controlled image manipulation. However, our model fails when errors exist in the condition semantic map. It would be an interesting future work to train the human parser and person image generation model jointly.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[3] Alex J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 8

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 7

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 4

[7] 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 *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3, 5, 6

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, 2014. 1, 2

[9] Aaron Hertzmann. Image analogies. *Proc Siggraph*, 2001. 8

[10] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Proc. Advances in Neural Information Processing Systems*, 2015. 4

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision*, 2016. 6

[14] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[15] 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. 2, 5

[16] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 5

[19] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Proc. Advances in Neural Information Processing Systems*, 2017. 1, 2, 3, 5, 6, 7

[20] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 6, 7

[21] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 4, 5, 6, 7

[22] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Proc. Advances in Neural Information Processing Systems*, 2016. 2

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Int'l Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proc. Advances in Neural Information Processing Systems*, 2016. 6

[25] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[26] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. Multistage adversarial losses for pose-based human image synthesis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[27] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 4, 5, 6, 7

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015. 4

[29] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[30] 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

[31] Wenqi Xian, Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[32] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proc. IEEE Int'l Conference on Computer Vision*, 2017. 2

[33] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE Int'l Conference on Computer Vision*, 2015. 5

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int'l Conference on Computer Vision*, 2017. 2, 4, 6

[35] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proc. IEEE Int'l Conference on Computer Vision*, 2017. 3