Unpaired Person Image Generation With Semantic Parsing Transformation

Sijie Song, *Student Member, IEEE*, Wei Zhang, Jiaying Liu, *Senior Member, IEEE*, Zongming Guo, *Member, IEEE*, and Tao Mei, *Fellow, IEEE*

Abstract—In this paper, we tackle the problem of pose-guided person image generation with unpaired data, which is a challenging problem due to non-rigid spatial deformation. Instead of learning a fixed mapping directly between human bodies as previous methods, we propose a new pathway to decompose a single fixed mapping into two subtasks, namely, semantic parsing transformation and appearance generation. First, to simplify the learning for non-rigid deformation, a semantic generative network is developed to transform semantic parsing maps between different poses. Second, guided by semantic parsing maps, we render the foreground and background image, respectively. A foreground generative network learns to synthesize semantic-aware textures, and another background generative network learns to predict missing background regions caused by pose changes. Third, we enable pseudo-label training with unpaired data, and demonstrate that end-to-end training of the overall network further refines the semantic map prediction and final results accordingly. Moreover, our method is generalizable to other person image generation tasks defined on semantic maps, e.g., clothing texture transfer, controlled image manipulation, and virtual try-on. Experimental results on DeepFashion and Market-1501 datasets demonstrate the superiority of our method, especially in keeping better body shapes and clothing attributes, as well as rendering structure-coherent backgrounds.

Index Terms—Image generation, semantic parsing transformation, appearance generation, fashion application

1 Introduction

Transferring a person image from one pose to another, which refers to pose-guided image generation in [1], has attracted great attention in recent years. The goal of this task is to change the pose of the person to a target one while keeping the appearance details at the same time. It is of great value in some fundamental computer vision tasks such as person re-identification and image/video manipulation. It also can be widely applied in art and fashion domains, benefiting applications such as on-demand movie production and fashion design.

The recent advent of deep learning and generative models [2] has provided powerful tools to achieve pose-guided image generation, and inspired many researchers in this area [1], [3], [4], [5], [6], [7], [8]. This problem is initially explored under a fully supervised setting [1], [5], [6], [7]. Though promising results have been presented, these methods require paired images (i.e., the same person in the same clothing but in different poses) in their training process. In order to address the data limitation and achieve more flexible generation, more recent works in this area focus on learning the mapping with unpaired data [3], [4], [8]. However, without supervision from ground truth images, the generated results from [3] are far from satisfactory due to the complexity in simultaneously modeling spatial and appearance transformations. Several works disentangle images into multiple factors, e.g., background and foreground [8], shape and appearance [4], [9], but ignoring non-rigid human-body deformations and clothing shapes can result in compromised quality of generated images. Another limitation of previous unpaired pose-guided image generation methods is that given a condition image, they mainly focus on rendering appearance-consistent foreground, without taking background synthesis into account. Therefore, the backgrounds in the generated results are less faithful to the condition images.

Therefore, the key challenges of this task with unpaired data are in the following aspects: (1) Due to the non-rigid nature of a human body, it is generally difficult to transform the spatially misaligned body-parts for convolution-based networks. (2) Clothing attributes, e.g., clothing types and textures, are difficult to preserve in the process of generation. However, these clothing attributes are important for human visual perception. (3) Pose changes would lead to missing regions in the background inherited from the condition image. It is troublesome to generate contextually-relevant background and stitch seamlessly with the foreground. (4) The lack of paired training data provides little clue in establishing effective training objectives.

In this paper, we seek a new pathway to address the above challenges. Rather than transforming the person image directly, we propose to introduce human semantic parsing as a bridge in the transformation. On one hand, translating between *semantic parsing* and *person image* (in

- *S. Song, J. Liu, and Z. Guo are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China. E-mail: {ssj940920, liujiaying, guozongming}@pku.edu.cn.*
- *W. Zhang and T. Mei are with the JD AI Research, Beijing 100105, China. E-mail: wzhang.cu@gmail.com, tmei@live.com.*

Manuscript received 30 Aug. 2019; revised 16 Mar. 2020; accepted 25 Apr. 2020. Date of publication 4 May 2020; date of current version 1 Oct. 2021. (Corresponding author: Jiaying Liu.) Recommended for acceptance by L. Sigal. Digital Object Identifier no. 10.1109/TPAMI.2020.2992105

both directions) has been extensively studied, where sophisticated tools [10], [11] are available. On the other hand, spatial deformation can be well handled by semantic parsing transformation, which is a much easier problem because the network does not need care about appearance information. Besides, semantic parsing naturally provides a foreground mask, which is important for seamlessly stitching the foreground and background in pose-guided image generation.

Specifically, our proposed model for person image generation with unpaired data comprises two modules: semantic parsing transformation and appearance generation. Semantic parsing transformation aims to transform semantic parsing between input and target poses with a semantic generative network. Based on the transformed semantic parsing, appearance generation module is designed to generate the foreground and background for the final output, respectively. A foreground generative network is proposed to synthesize semantic-aware textures on the transformed parsing. And a background generative network is developed to predict missing background regions caused by pose changes, so that structurally-coherent background can be rendered according to the condition image. Without paired supervision, training the proposed network is intractable as different modules are highly coupled with each other. Therefore, we are motivated to seek a divide-and-conquer training strategy. Each part of the network is first independently trained and then jointly optimized with each other. For semantic parsing transformation, we create pseudo labels to guide the network training. For appearance generation, we adopt cycle consistency to overcome the absence of ground truth images. In addition, we propose a semantic-aware style loss, encouraging the foreground generative network to build a mapping between corresponding semantic regions, thus clothing attributes can be well-preserved by rich semantic parsing. In order to generate natural and coherent background, we train the background generative network with auxiliary images, and use an iterative optimization procedure to adapt both auxiliary and original images.

Moreover, we are inspired to apply the appearance generative network on conditional image generation tasks, thanks to the mapping between corresponding semantic regions. Guided by the semantic map, we can transfer clothing textures of two person images, or control the image generation by editing the semantic map manually. Meanwhile, we are able to apply our model on virtual try-on, fitting new clothes or styled textures to the target person.

We summarize our main contributions as follows:

- To address the challenging problem of person image generation with unpaired data, we propose to decompose it into two subtasks, namely, semantic parsing transformation (H_S) and appearance generation (H_A).
- A delicate training schema is designed to carefully optimize H_S and H_A in a divide-and-conquer manner. We enable a pseudo-label training process with unpaired data, and demonstrate that end-to-end training of the network enables better semantic map prediction, and then helps improve the final results.
- For appearance generation, we consider generating foreground and background separately. We develop a background generative network to predict the missing background regions caused by pose changes and generate realistic-looking outputs.
- Our model is superior in keeping clothing attributes, rendering better body shape, and retaining a coherent background from the condition image. It is also generalizable to other conditional image generation applications, including clothing texture transfer, controlled image manipulation, and virtual try-on.

A preliminary version of our work has been presented in [12]. In this journal article, our work is improved from the following aspects: (1) We improve the proposed person image generation method by taking background generation into account. We aim to inherit the background information from the condition image in the final output, which is overlooked by the previous methods [1], [3], [4], [5], [8]. (2) We provide more details of our proposed method and present more extensive analysis of our model. We elaborate on the effectiveness of each component in semantic parsing transformation and appearance generation, including the impact of person representation and different loss function terms, as well as the effect of various background generation schemes. (3) We explore another application scenario of virtual try-on, to further inspire the fashion community. It is worth noting that our model not only fits new clothes but also transfers texture styles on the target person.

The paper is structured as follows. In Section 2, we discuss related work on image generation. In Section 3, we introduce the proposed model for person image generation with unpaired data. Subsequently, we present our experiments and extensive analysis on two datasets (i.e., DeepFashion [13], Market-1501 [14]) in Section 4. The concluding remarks are given in Section 5.

2 Related Work

2.1 Image Generation

The generative models, such as variational autoencoders (VAEs) [15], [16] and generative adversarial networks (GANs) [2], have been significantly improved in the past few years. Their ability in synthesizing realistic-looking and natural images has led the progress in image generation [17], [18], [19], [20], [21], [22]. VAE-based methods learn the mappings between domains by optimizing negative log-likelihood of the training data [21], [23], [24]. They are easy to train but usually produce very blurry images that lack details. GANs, on the other hand, optimize a min-max objective with a generator and a discriminator. Though facing challenges in training stability, GANs tend to generate more realistic images with sharper edges, and many efforts have been made to effectively enable a stable training process [25], [26], [27]. Therefore, GAN-based image generation also attracts much attention.

For GAN-based image generation methods, there are mainly two branches: supervised methods and unsupervised methods. With paired training data in the supervised setting, pix2pix [10] achieves image to image translation, which is essentially a domain transfer problem, by building a conditional GAN. More recent efforts [17], [18] have been dedicated to generating photo-realistic images in high-resolution by generating multi-scale images progressively. For the

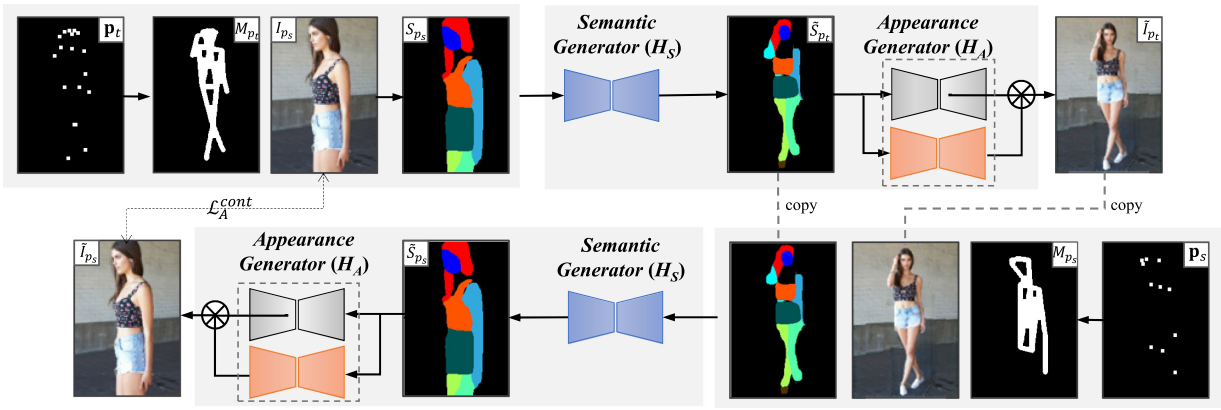


Fig. 1. Our overall framework for unpaired person image generation. We decompose the mapping into semantic parsing transformation (H_S) and appearance generation (H_A). Note that H_A consists of two streams: a foreground generative network (in grey) and a background generative network (in orange) to render foreground and background, respectively. H_S generates the semantic map $\tilde{S}_{p_t}$ under the target pose $\mathbf{p}_t$, and H_A further generates the output $\tilde{I}_{p_t}$ guided by $\tilde{S}_{p_t}$. Then we achieve the cycle to get the recovered input $\tilde{I}_{p_s}$. For simplicity, we omit some inputs of the generator. The detailed illustration of H_S and H_A can be found in Figs. 2 and 4, respectively.

unsupervised setting, the works in [28], [29], [30] employ reconstruction consistency to learn cross-domain mapping. However, the development and application of these unsupervised methods are mainly for appearance generation in spatially aligned tasks. With unpaired training data, we aim to learn a mapping to simultaneously deal with spatial non-rigid deformation and appearance generation.

2.2 Pose-Guided Person Image Generation

For the problem of pose-guided person image generation, the two-stage network PG^2 [1] is one of the early attempts. It first coarsely generates an output under the target pose, and then refines the result with finer details. To better model appearance and shape, Siarohin *et al.* [5] proposed to transform high-level features for each body part with deformable skips. The similar idea is also adopted by [7], [31], in which segmentation masks of body parts are employed to guide the image generation. A more recent work [32] builds correspondence between the source and target images by texture coordinate estimation and inpainting. However, the models in [1], [5], [7], [32] are trained using paired data.

To overcome the limitation, the work in [3] presents a fully unsupervised GAN based on [28], [33]. Dominik *et al.* [9] proposed an approach with a two-stream auto-encoder for unsupervised learning of shape and appearance. Other works [4], [8] tackle the problem with unpaired data by modeling data distribution and sampling from feature spaces. Generated from highly-compressed features, the results of these methods are less consistent with the appearance of condition images. Instead, we leverage semantic information as guidance in body shape generation and texture synthesis for the final output.

In most of the previous pose-guided image generation methods [1], [3], [4], [5], [8], background generation is always overlooked. However, changing the human pose of the input image while maintaining the background is crucial in generating temporally coherent videos. While both [6] and [7] employ an additional module to generate images conditioned upon the backgrounds from the input images, they learn to render the missing regions caused by pose changes with supervision from ground truth images. In our work, we propose a model to synthesize coherent background images with unpaired data, which is more challenging for the model to infer context-aware textures.

2.3 Semantic Parsing for Image Generation

Semantic maps provide valuable prior for image generation. Starting from the pixel-wise semantic map, pix2pix [10] made a breakthrough in structure-conditional image translation. Then semantic maps are widely employed in visual manipulation [34], [35], which allows semantic control over the process in image generation. Besides, semantic maps can also serve as an intermediate representation between conditional inputs and output images. The condition inputs can be texts [36], scene graphs [37] or human skeletons [38]. These works first infer the semantic maps, and then generate an image with an image generator. The results in [36], [37] show the effectiveness of predicting scene layout (semantic map) for text-to-image translation. It is illustrated that by conditioning on estimated layouts or semantic maps, more semantically meaningful images can be generated. More recent works in [38], [39], [40] apply the idea in person image generation by inferring human parsing, and benefit from semantic information of human body structures. However, these models have to be trained with ground truth supervision to predict scene layouts or semantic maps. In contrast, our model learns semantic map prediction with unpaired data. We further demonstrate that the prediction for the semantic map can be improved with end-to-end training.

3 The Proposed Method

Given a target pose $\mathbf{p}_t$ and a condition 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: $< I_{p_s}, \mathbf{p}_t > \rightarrow \tilde{I}_{p_t}$.

During the most practical scenario, we are working under an unpaired 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 problem of unpaired person image generation, our key idea is to introduce human semantic parsing to decompose it into two modules: *semantic parsing transformation* and *appearance generation*. Fig. 1 shows our overall framework. The semantic

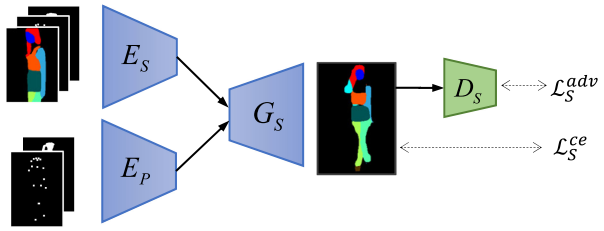


Fig. 2. Semantic parsing transformation module. The semantic generator H_S consists of a semantic map encoder E_S, a pose encoder E_P, and a semantic map generator G_S, which can be formulated as $H_S = G_S \circ (E_S, E_P)$.

parsing transformation module aims to generate a semantic map under the target pose, providing valuable prior-knowledge for the clothing attributes and human body shape. Guided by the condition image and the predicted semantic map, the appearance generation module then synthesizes textures for the foreground and background of the final output, respectively.

In the following, we first introduce person representation (Section 3.1), which is the input of our framework. We then describe each module in detail from the perspective of independent training (Sections 3.2 and 3.3). Finally, we illustrate our training strategy to jointly optimize the two modules (Section 3.4).

3.1 Person Representation

The input of our model includes the condition image $I_{p_s} \in \mathbb{R}^{3\times H\times W}$, the source pose $\mathbf{p}_s$, and the target pose $\mathbf{p}_t$. Besides, our framework 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, poses are represented as probability heat maps, i.e., $\mathbf{p}_s, \mathbf{p}_t \in \mathbb{R}^{k\times H\times W} (k = 18)$. We extract the semantic map S_{p_s} with an off-the-shelf human parser [11]. S_{p_s} is encoded as a pixel-level one-hot vector, i.e., $S_{p_s} \in \{0,1\}^{L\times H\times W}$, and L indicates the total number of semantic labels. We adopt the same definition in [1] for the pose masks M_{p_s} and M_{p_t}, as priors for pose joint connection in image generation.

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 condition semantic map S_{p_s}. It is achieved by the semantic generative network based on U-Net [41]. As shown in Fig. 2, 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 semantic 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. To generate the semantic label for each pixel, we employ *softmax* activation function as [42] at the end of G_S. The predicted semantic map $\tilde{S}_{p_t}$ conditioned on S_{p_s} and $\mathbf{p}_t$ can be formulated as $\tilde{S}_{p_t} = G_S\big(E_S(S_{p_s}, \mathbf{p}_s, M_{p_s}), E_P(\mathbf{p}_t, M_{p_t})\big)$. 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. As clothing textures are not associated with semantic maps, people in different clothing appearances may share similar semantic maps. Therefore, we can create semantic map pairs from training images to facilitate model training. For a given S_{p_s}, we search a semantic map $S_{p_t^*}$, which is under a different pose 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 such a search problem. As shown in Fig. 3, we divide the human body into ten rigid body parts following [5], 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

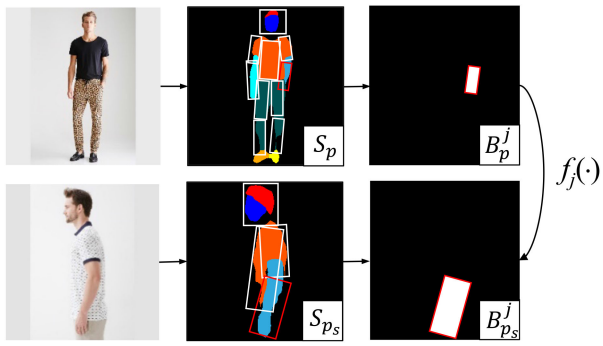


Fig. 3. In the process of pseudo label searching, the human bodies are decomposed into rigid parts and aligned by affine transformations.

$$S_{p_t^*} = \arg\min_{S_p} \frac{1}{n} \sum_{j=j_1}^{j_n} \frac{1}{|B_{p_s}^j|} ||B_{p_s}^j \otimes S_{p_s} - f_j(B_p^j \otimes S_p)||_2^2, \tag{1}$$

where $\{j_1, \ldots, j_n\}$ denote the binary mask indexes that are valid both for S_p and S_{p_s}. $\otimes$ is the element-wise multiplication operator. We align the two body parts with an affine transformation $f_j(\cdot)$, which can be calculated by minimizing least-squares error according to four corners of corresponding binary masks

$$\min_{f_j(\cdot)} ||f_j(B_p^j) - B_{p_s}^j||_2^2. \tag{2}$$

Note that during pseudo label generation, pairs sharing very similar poses are excluded. In practice, we randomly choose N ($N = 500$) images from the training set as pseudo label candidates. Then we perform the pseudo label generation described above, and choose the pseudo label from candidates by solving Eq. (1). On the one hand, by randomly choosing pseudo label candidates, we avoid the situation that pseudo label generation finds ground-truth labels. On the other hand, it accelerates the searching process, since it would be very slow to match the whole training set with the given semantic map individually.

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

$$\mathcal{L}_S^{ce} = -||S_{p_t^*} \otimes log(\tilde{S}_{p_t^*}) \otimes (1 + M_{p_t^*})||_1. \tag{3}$$

Adversarial Loss. The adversarial loss $\mathcal{L}_S^{adv}$ is employed with a discriminator D_S to help G_S generate semantic maps that are visually similar to realistic ones.

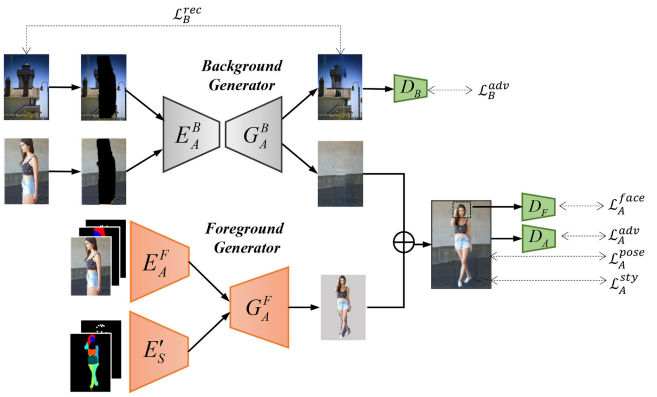


Fig. 4. Appearance generation. We consider two streams: foreground and background generation. The appearance generator is formulated as $H_A = (G_A^F \circ (E_A^F, E_S')) \oplus (G_A^B \circ E_A^B)$.

$$\mathcal{L}_S^{adv} = \mathcal{L}^{adv}(H_S, D_S, S_{p_t^*}, \tilde{S}_{p_t^*}), \tag{4}$$

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 the output of G.

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

$$\mathcal{L}_S^{total} = \mathcal{L}_S^{adv} + \lambda^{ce}\mathcal{L}_S^{ce}. \tag{5}$$

3.3 Appearance Generation (H_A)

In this module, we synthesize textures for the output image $\tilde{I}_{p_t} \in \mathbb{R}^{3\times H\times W}$, guided by the condition image I_{p_s} and the predicted semantic map $\tilde{S}_{p_t}$ from our semantic parsing transformation module. Inspired by [43], we consider the appearance generation for foreground $\tilde{I}_{p_t}^F$ and background $\tilde{I}_{p_t}^B$, respectively. Then, the generation of image $\tilde{I}_{p_t}$ can be factorized as

$$\tilde{I}_{p_t} = \tilde{I}_{p_t}^F \otimes \Omega(\tilde{S}_{p_t}) + \tilde{I}_{p_t}^B \otimes (\mathbf{1} - \Omega(\tilde{S}_{p_t})), \tag{6}$$

where $\otimes$ indicates element-wise multiplication, $\Omega(\cdot)$ is to obtain the binary foreground mask with the semantic map. Accordingly, our appearance generation module consists of two main pieces, i.e., a foreground generative network and a background generative network, as shown in Fig. 4.

Foreground Generation. The foreground generative network comprises an appearance encoder E_A^F to extract the foreground appearance of condition image I_{p_s}, a semantic map encoder E_S' to encode the predicted semantic map $\tilde{S}_{p_t}$, and a foreground generator G_A^F. Different from the semantic generative network, deformable skips in [5] are adopted to better handle spatial deformation. We obtain the foreground of the output image by $\tilde{I}_{p_t}^F = G_A^F(E_A^F(I_{p_s}, S_{p_s}, \mathbf{p}_s), E_S'(\tilde{S}_{p_t}, \mathbf{p}_t))$.

Background Generation. To keep the background of the input image, we adopt a basic U-Net [41] as the background generative network, which consists of a background encoder E_A^B and a background generator G_A^B. The background of the output image is obtained by $\tilde{I}_{p_t}^B = G_A^B\left(E_A^B(I_{p_s} \otimes \Omega(S_{p_s}))\right)$.

The generated foreground $\tilde{I}_{p_t}^F$ and background $\tilde{I}_{p_t}^B$ are then combined to get the final output image $\tilde{I}_{p_t}$ with Eq. (6). We formulate the overall appearance generation process as $H_A = \left(G_A^F \circ (E_A^F, E_S')\right) \oplus \left(G_A^B \circ E_A^B\right)$. Without the supervision from ground truth I_{p_t}, the appearance generation module is trained based on cycle consistency as [3], [28], in which H_A should be able to map back I_{p_s} with the generated $\tilde{I}_{p_t}$ and $\mathbf{p}_s$. In the process of mapping back, the mapped-back image is denoted as $\tilde{I}_{p_s}$, and the predicted segmentation map is represented as $\tilde{S}_{p_s}$.

Adversarial Loss. We first introduce the discriminator D_A to distinguish between generated and realistic images, which leads to an adversarial loss $\mathcal{L}_A^{adv}$

$$\mathcal{L}_A^{adv} = \mathcal{L}^{adv}(H_A, D_A, I_{p_s}, \tilde{I}_{p_t}) + \mathcal{L}^{adv}(H_A, D_A, I_{p_s}, \tilde{I}_{p_s}). \tag{7}$$

Pose Loss. To generate images faithful to the target pose, we use a pose loss $\mathcal{L}_A^{pose}$ with a pose detector $\mathcal{P}$ as [3]

$$\mathcal{L}_A^{pose} = ||\mathcal{P}(\tilde{I}_{p_t}) - \mathbf{p}_t||_2^2 + ||\mathcal{P}(\tilde{I}_{p_s}) - \mathbf{p}_s||_2^2. \tag{8}$$

Content Loss. We also employ a content loss $\mathcal{L}_A^{cont}$ to ensure cycle consistency

$$\mathcal{L}_A^{cont} = \sum_{i=0}^{2} ||\alpha_i \Lambda_i(\tilde{I}_{p_s}) - \Lambda(I_{p_s})||_2^2, \tag{9}$$

where $\Lambda_i(I)$ is the feature map of image I of the ith layer in VGG16 model [44] pretrained on ImageNet. In practice, we use $conv1\_2$ and $conv2\_1$ for $i \geq 1$ and RGB pixels for $i = 0$.

Style Loss. Since I_{p_s} and $\tilde{I}_{p_t}$ are spatially misaligned, it is challenging to transfer textures and color information correctly without any constraints. The work in [3] tackled this issue using patch-style loss. It enforces textures around corresponding pose joints in I_{p_s} and $\tilde{I}_{p_t}$ to be 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. To address the above issues, we design a semantic-aware style loss to well retain the style, thanks to the guidance provided by semantic maps. By enforcing the style consistency among I_{p_s}, $\tilde{I}_{p_t}$ and $\tilde{I}_{p_s}$, we define the semantic-aware style loss as

$$\mathcal{L}_A^{sty} = \mathcal{L}^{sty}(I_{p_s}, \tilde{I}_{p_t}, S_{p_s}, \tilde{S}_{p_t}) + \mathcal{L}^{sty}(\tilde{I}_{p_t}, \tilde{I}_{p_s}, \tilde{S}_{p_t}, \tilde{S}_{p_s}), \tag{10}$$

where

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

$\mathcal{G}(\cdot)$ denotes the function for Gram matrix [45], $\Psi_l(S)$ denotes the downsampled binary map from S, indicating pixels that belong to the lth semantic label.

Face Loss. To generate more natural-looking faces, a face loss $\mathcal{L}_A^{face}$ is added with the discriminator D_F

$$\begin{aligned}\mathcal{L}_A^{face} &= \mathcal{L}^{adv}(H_A, D_F, \mathcal{F}(I_{p_s}), \mathcal{F}(\tilde{I}_{p_t}))\\ &\quad + \mathcal{L}^{adv}(H_A, D_F, \mathcal{F}(I_{p_s}), \mathcal{F}(\tilde{I}_{p_s})),\end{aligned} \tag{11}$$

where $\mathcal{F}(I)$ represents the face extraction guided by facial joints. In our experiments, we perform it using a non-parametric spatial transform network [46].

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

$$\mathcal{L}_A^{total} = \mathcal{L}_A^{adv} + \lambda^{pose}\mathcal{L}_A^{pose} + \lambda^{cont}\mathcal{L}_A^{cont} + \lambda^{sty}\mathcal{L}_A^{sty} + \mathcal{L}_A^{face}. \tag{12}$$

For now, we take an overall consideration of training for foreground and background generation. One of the challenges is that the changes in human poses would lead to missing regions in the background for the output image. To generate natural images, it is expected that the background generative network would achieve inpainting according to the spatial context. However, the lack of supervision from ground truth background images makes it hard for the network to correctly render the missing regions, especially when there are complex textures in the background. To tackle this issue, we introduce an auxiliary dataset (i.e., Place2 [47]) to help train the background generative network.

As shown in Fig. 4, guided by the foreground mask, we remove the corresponding regions both from the auxiliary image and the condition image. For auxiliary images, we train the background generative network by regressing to the ground truth with an L2 distance [48]

$$\mathcal{L}_B^{rec} = ||I_{aux} - G_A^B\big(E_A^B(I_{aux} \otimes \Omega(S_{p_s}))\big)||_2^2, \tag{13}$$

where I_{aux} is the ground truth image from the auxiliary dataset. Adversarial loss $\mathcal{L}_B^{adv}$ is also employed to make prediction realistic and coherent to the context. Since there is a domain gap between the auxiliary and condition images, we use another discriminator D_B to achieve the adversarial learning. For a more stable training, we adopt an iterative optimization procedure as illustrated in Algorithm 1.

Algorithm 1. Iterative Training for Appearance Generation

Input: $\{I_{p_s}^i, S_{p_s}^i, \mathbf{p}_s^i, \tilde{S}_{p_t}^i, \mathbf{p}_t^i\}_{i=1}^N$, $\{I_{aux}^i\}_{i=1}^{N^{aux}}$, number of training iterations K.

1: **for** $t = 1, \ldots, K$ **do**
2: sample $\{I_{p_s}^i, S_{p_s}^i, \mathbf{p}_s^i, \tilde{S}_{p_t}^i, \mathbf{p}_t^i\}$.
3: forward H_A to perform pose guided image generation.
4: update H_A by minimizing $\mathcal{L}_A^{total}$.
5: update D_F, D_A by maximizing $\mathcal{L}_A^{face} + \mathcal{L}_A^{adv}$.
6: sample $\{I_{aux}^i, S_{p_s}^i\}$.
7: forward E_A^B, G_A^B to perform image inpainting.
8: update E_A^B, G_A^B by minimizing $\alpha\mathcal{L}_B^{rec} + \beta\mathcal{L}_B^{adv}$ ($\alpha = 1$, $\beta = 10$).
9: update D_B by maximizing $\mathcal{L}_B^{adv}$.
10: **end for**

Output: H_A.

Some could argue that training an independent network for inpainting with an auxiliary dataset could be enough for background generation. We conducted such experiments and found that the domain gap between the auxiliary and original datasets leads to artifacts on the generated images. Our training scheme, however, can adapt to both domains and generate more pleasant results.

3.4 End-to-End Training

As the contour and shape of the final generated images are defined by human parsing, the quality of semantic map prediction from semantic parsing transformation decides the visual results of appearance generation. However, if the two modules (H_S and H_A) are trained independently, two reasons may cause instability.

(a) Searching error (b) Parsing error

Fig. 5. Potential errors in the searched semantic map pair that might harm semantic parsing transformation.

- Searching error. As shown in Fig. 5a, the searched semantic maps are not accurate enough. For example, the sleeve lengths are not consistent in the searched pairs (indicated by the white arrows).
- Parsing error. As shown in Fig. 5b, the semantic maps extracted by the human parser are inaccurate, because ground truth semantic labels are not available to finetune the human parser.

For refining the semantic parsing transformation module, the overall framework in Fig. 1 is trained in an end-to-end manner. Our training scheme is shown in Algorithm 2.

Algorithm 2. End-to-End Training for the Overall Framework

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 $\{I_{p_s}^i, S_{p_s}^i, \mathbf{p}_s^i, \mathbf{p}_t^i\}_{i=1}^N$, $\{I_{aux}^i\}_{i=1}^{N^{aux}}$ and $\{H_S, D_S\}$ fixed, train $\{H_A, D_A, D_F, D_B\}$ with the iterative optimization procedure in Algorithm 1.
//Joint optimization
4: With E_A^B and G_A^B fixed, train $\{H_S, D_S, H_A, D_A, D_F\}$ jointly with $\mathcal{L}_A^{total}$, using $\{I_{p_s}^i, S_{p_s}^i, \mathbf{p}_s^i, \mathbf{p}_t^i\}_{i=1}^N$.

Output: H_S, H_A.

4 Experiments

We evaluate our proposed model both qualitatively and quantitatively in this section. Moreover, we give an extensive ablation study to better analyze each component of our model. In the end, we show various applications of our proposed model to inspire the fashion community.

4.1 Datasets and Settings

DeepFashion [13]. The subset of DeepFashion, i.e., *In-shop Clothes Retrieval Benchmark* comprises a great number of clothing images in different poses and appearances. The images have a resolution of 256×256. To evaluate our foreground and background generation respectively, we adopt two settings for the DeepFashion dataset as follows.

- DeepFashion w/o b.g: In this setting, we evaluate foreground generation. As paired data are not required by our method, 37,258 images are randomly

selected for training and 12,000 images for testing. The images for testing are with easy backgrounds.

- DeepFashion w/ b.g: In this setting, we evaluate foreground generation and background generation. As DeepFashion w/o b.g., we use 37,258 images from the DeepFashion dataset. To help background generation, we adopt the validation set from Place2 [47], which consists of 36,500 images, as auxiliary data. We manually select 118 images with complex backgrounds from DeepFashion test set to evaluate the performance.

Market-1501 [14]. The images in this dataset were captured from different viewpoints. There are 32,886 images in total with a resolution of 128×64. We adopt the same protocol defined in [14] to obtain data splits, and select 12,000 pairs of images for testing as [5].

Implementation Details. For the person representation, we extract 2D poses with OpenPose [49], and obtain condition semantic maps with a state-of-the-art human parser [11]. The semantic labels originally defined in [11] are integrated into ten categories ($L = 10$), including background, face, hair, upper clothes, pants, skirt, left/right arm, left/right leg. For the DeepFashion dataset, we perform the joint learning on a resolution of 128×128 to refine semantic map prediction. Then the predicted semantic maps are upsampled and we train images in 256×256 using progressive training strategies introduced in [17]. For Market-1501, the model is directly trained and tested on 128×64. In addition, as the images in Market-1501 are in a low resolution with blurry faces, we omit $\mathcal{L}_A^{face}$ for appearance generation on this dataset. For the hyper-parameters, λ^{pose}, λ^{cont} are set as 700, 0.03 for DeepFashion and 1, 0.003 for Market-1501, respectively. λ^{sty} is 1 for all experiments. ADAM optimizer [50] is employed to train the network with a learning rate 0.0002 ($\beta_1 = 0.5$ and $\beta_2 = 0.999$). The batch size is set as 4 for DeepFashion and 16 for Market-1501, respectively. We use four P40 GPUs to train our model. For training images in the resolution of 128×128 (256×256), it costs about 1 h$\sim$2 h (5 h$\sim$6 h) on each epoch. In total, we spent 5 days training DeepFashion and less than one day training Market-1501.

4.2 Comparison With State-of-the-Art Methods

In this subsection, we compare our model with four state-of-the-art methods: PG^2 [1], Def-GAN [5], UPIS [3], and V-UNet [4].[1] PG^2 [1] and Def-GAN [5] require paired training data, while UPIS [3] and V-UNet [4] do not. Note that for a fair comparison, we test images from DeepFashion with easy backgrounds (DeepFashion w/o b.g.) and complex backgrounds (DeepFashion w/ b.g.), respectively. For Market-1501, the backgrounds are cluttered and blurry, explicit background generation is not performed.

Qualitative Comparison. In Figs. 6 and 7, we present the qualitative comparison on the DeepFashion dataset. Our model generates more photo-realistic results with fewer artifacts and higher visual quality. Fig. 6 shows the results conditioned on the images with easy backgrounds to give a comparison in the foreground generation. Our method is

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

Fig. 6. Example results of images for DeepFashion (w/o b.g.) by different methods (PG^2 [1], Def-GAN [5], UPIS [3], and V-UNet [4]). Our model better keeps clothing attributes (e.g., textures, clothing types).

superior especially in keeping the clothing attributes, including textures and clothing type (the last row). Fig. 7 shows the results that conditioned on the images with diverse and complex backgrounds. The overlook of background generation in state-of-the-art methods [1], [3], [4], [5] leads to blurry backgrounds with many texture details missing. In contrast, our model not only generates pleasing foregrounds, but also keeps the backgrounds from the condition images. The results also demonstrate that our model is able to inpaint missing background regions caused by pose changes with context-aware textures.Fig. 8 shows the qualitative results on the Market-1501 dataset. It can be seen that our method better shapes the legs and arms. Even without using background generative network to explicitly reconstruct backgrounds, our model also successfully retains background information from condition images.

Quantitative Results. Table 1 shows the quantitative evaluation in the metrics of Inception Score (IS) [51], Structural SIMilarity (SSIM) [52] and Fréchet Inception Distance (FID) [53]. When computing IS and SSIM for Market-1501,

Fig. 7. Example results of images for DeepFashion (w/ b.g.) by different methods (PG^2 [1], Def-GAN [5], UPIS [3], and V-UNet [4]). Our model is able to retain the backgrounds from condition images and generate natural results.

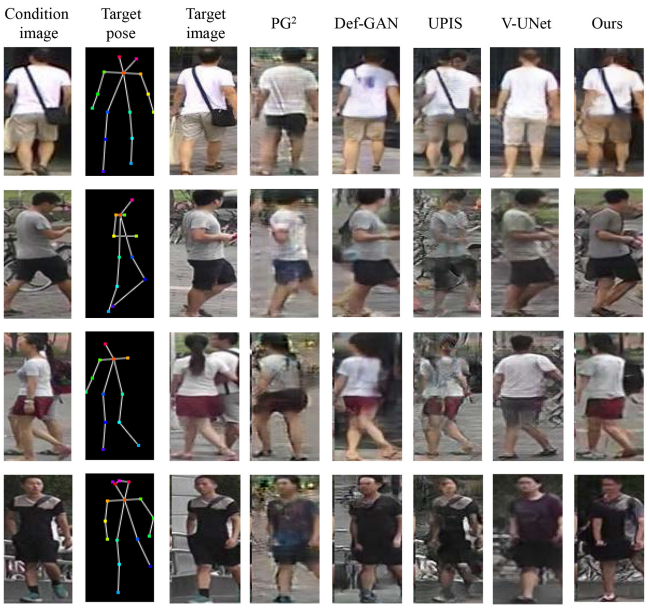


Fig. 8. Example results by different methods (PG^2 [1], Def-GAN [5], UPIS [3], and V-UNet [4]) on Market-1501. Our model generates better body shapes.

we also employ mask-IS and mask-SSIM as [1] to exclude the background area, since there is a large difference between the backgrounds of condition and target images. Besides, the requirements of training data are marked for each method in Table 1 for a fair comparison.

In all the settings, our proposed model achieves the best IS and FID, even compared with supervised methods. It is consistent with better body shapes and more realistic details in our results. For SSIM score, our results are slightly lower than other methods in DeepFashion (w/o background) and Market-1501, and comparable with state-of-the-art methods in DeepFashion (w/ background). It can be explained by the fact that blurry images always get higher SSIM but being less photo-realistic, as also observed in [1], [8], [54], [55].

User Study. To give a more comprehensive comparison, we further implement a user study to evaluate our pose-guided person image generation results compared with other state-of-the-arts. For each dataset, we perform pairwise A/B tests to 35 volunteers, and everyone is given 250 pairs that are randomly selected from the results. In each pair, the images are in random order, one of which is our result while the other is from the compared method. Volunteers are asked to select the better one without time limit, considering: (1) correctly change the pose of the person in the condition images, (2) correctly preserve the clothing attributes (e.g., textures, colors, clothing types) from the condition images to the target images, (3) correctly retain the background information from the condition images, (4) natural and photo-realistic visual quality.

TABLE 2
User Study Results on Pose-Guided Person Image Generation

	Deep Fashion (w/o b.g.)		Deep Fashion (w/ b.g.)		Market-1501	
	mean	var	mean	var	mean	var
PG^2 [1]	92.20%	0.007	95.83%	0.004	78.55%	0.019
Def-GAN [5]	64.61%	0.036	67.13%	0.031	67.23%	0.022
UPIS [3]	97.90%	0.001	97.92%	0.001	86.17%	0.012
V-UNet [4]	68.55%	0.007	70.70%	0.030	73.31%	0.046

The mean values indicate the percentages of volunteers that select our method as better results in pairwise comparisons.

Table 2 shows our user-study results for mean and variance values. The mean values indicate the percentages of volunteers that select our method as better results in pairwise comparisons. For example, about 92.20 percent volunteers think our method generates better images than PG^2 [1]. The variance values indicate how volunteers think differently for the given pairs. The results in Table 2 illustrate that our method effectively generates images with more pleasant visual quality than state-of-the-art methods.

4.3 Ablation Study

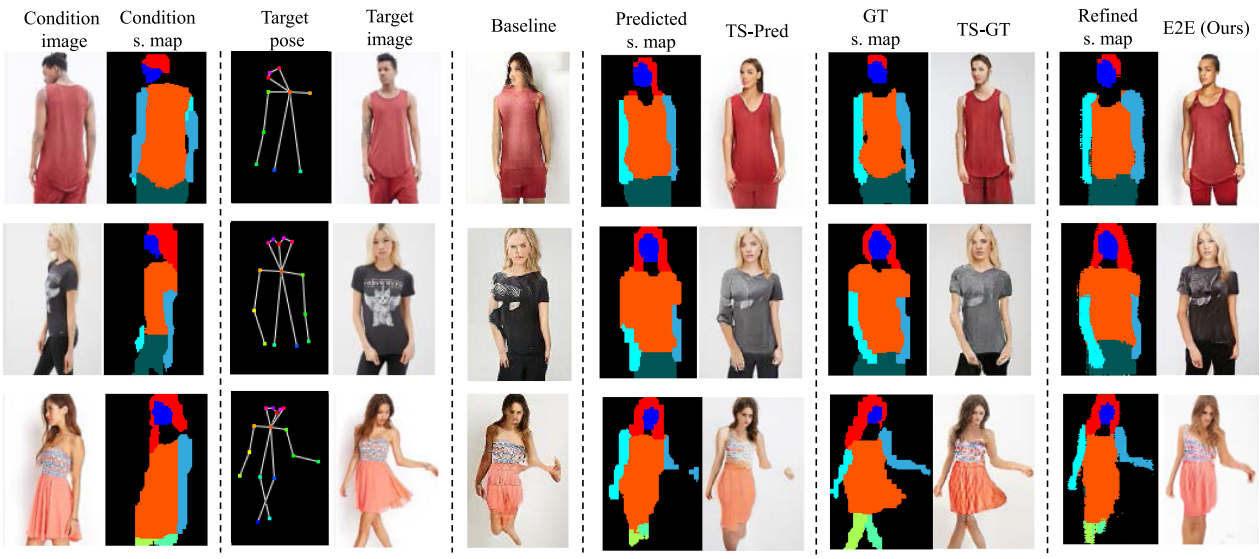
In this subsection, we first evaluate the introduction of semantic parsing, then elaborate on the effectiveness of each component in semantic parsing transformation and appearance generation to better understand our model.

To evaluate the introduction of semantic information in unpaired person image generation, we design the following experiments with different configurations.

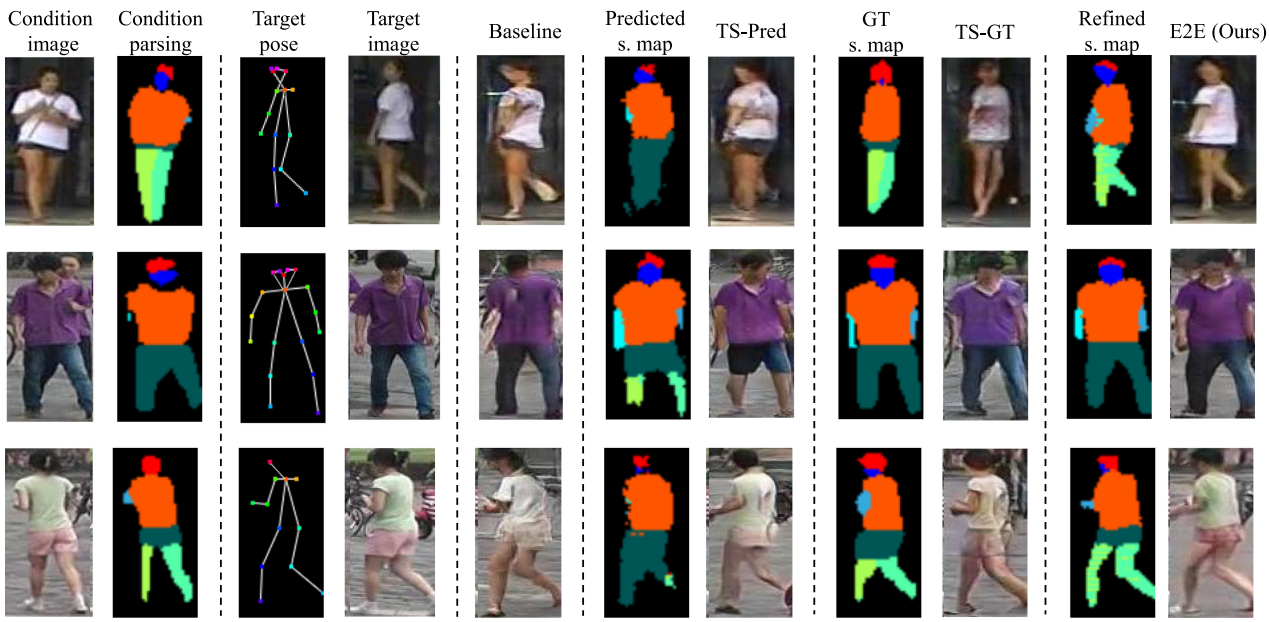
- Baseline: The architecture of our baseline model is the same as the appearance generation module. It directly learns the mapping between input and output images without the introduction of semantic information. Mask-style loss, which employs body part masks instead of semantic maps in Eq. (10), is used to keep the style on the final result.
- TS-Pred: We train the semantic parsing transformation and appearance generation independently in two-stage. The predicted semantic maps are fed into

TABLE 1
Quantitative Results on DeepFashion and Market-1501 Datasets (*Based on Our Implementation)

		DeepFashion (w/o b.g.)			DeepFashion (w/ b.g.)			Market-1501				
Models	Paired data	IS ↑	SSIM ↑	FID ↓	IS ↑	SSIM ↑	FID ↓	IS ↑	SSIM ↑	mask-IS ↑	mask-SSIM ↑	FID ↓
PG^2 [1]	Y	3.090	0.762	47.621	2.286	0.651	78.967	3.460	0.253	3.435	0.792	142.731
Def-GAN [5]	Y	3.439	0.756	18.672	2.191	0.693	55.752	3.185	0.290	3.502	**0.805**	78.305
V-Unet [4]	N	3.087	**0.786**	26.043	2.513	0.637	61.470	3.214	**0.353**	–	–	211.710
UPIS [3]	N	2.971	0.747	23.364	2.162	0.566	63.529	3.431*	0.151*	3.485*	0.742*	70.008
Ours	N	**3.441**	0.736	**12.225**	**2.633**	**0.699**	**55.259**	**3.499**	0.203	**3.680**	0.758	**56.442**



(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.)

Fig. 9. Ablation studies on semantic parsing transformation.

the appearance generation module to generate the final results.

- TS-GT: We train the semantic parsing transformation and appearance generation independently in two-stage. The semantic maps extracted from target images are regarded as ground truth semantic labels, and then fed into the appearance generation module to generate the final result.
- E2E (Ours): The overall framework is optimized jointly by end-to-end training.

We show the intermediate semantic maps and corresponding generated images in Fig. 9. It is difficult for the network to handle the appearance and shape simultaneously without the guidance from semantic maps. However, the results with semantic parsing transformation outperform the baseline consistently. It is observed that the image quality drops directly due to the errors in the predicted semantic maps when trained in two-stage, but the semantic map prediction can be refined in end-to-end training. For instance, our model is able to well preserve the haircut and sleeves length in Fig. 9a. We further present the quantitative results of different configurations in Table 3. For DeepFashion, we obtain comparable IS, SSIM, and FID scores in end-to-end training strategy (E2E) with those using ground truth semantic maps (TS-GT). For Market-1501, we obtain even better results from E2E than TS-GT, mainly because the human parser [11] does not work well on low-resolution images and there are many errors in the parsing results, as the first row in Fig. 9b.

TABLE 3
Quantitative Results Under Different Configurations on DeepFashion (w/o b.g.) and Market-1501 Datasets

	DeepFashion			Market-1501				
Models	IS ↑	SSIM ↑	FID ↓	IS ↑	SSIM ↑	mask-IS↑	mask-SSIM ↑	FID ↓
Baseline	3.140	0.698	15.443	2.776	0.157	2.814	0.714	75.436
TS-Pred	3.201	0.724	13.881	3.462	0.180	3.546	0.740	66.858
TS-GT	3.350	**0.740**	12.386	3.472	0.200	3.675	0.749	58.877
E2E	**3.441**	0.736	**12.225**	**3.499**	**0.203**	**3.680**	**0.758**	**56.442**

4.3.1 Analysis of Semantic Parsing Transformation

We now give more detailed analysis of semantic parsing transformation. The experiments are conducted with the semantic generative network trained in two-stage. We explore the effectiveness of person representation and loss functions.

• *Person Representation.* We first investigate the effectiveness of person representation in semantic parsing transformation. It is found that pose masks play a key role in generating a smooth and natural semantic map under the target pose. As shown in Fig. 10a, after removing pose masks from the input, the semantic generative network tends to generate breaking limbs (arms in the 1st and 3rd rows) or unnatural body shapes (arm in the 2nd row). We also explore the effect of pose heat maps by removing them from the input and keeping pose masks. The results can be viewed in Fig. 10b. Without pose heat maps, though pose masks have provided joint positions implicitly, the model fails to generate natural results, due to the lack of semantic information represented through different channels in the pose heat maps. However, together with the joint connection prior from pose masks and explicit semantic information from pose heat maps, we obtain continuous and natural semantic maps (see *Ours* in Fig. 10).

• *Pseudo Label Generation.* With our pseudo label generation strategy, we explore if any semantic maps find their ground-truths as pseudo labels, and show how it affects the semantic parsing transformation performance. We experiment with different $N(N = 100, 500, 1000)$, which is the number of candidate images, to obtain different sets and train the semantic generator. We first calculate how many semantic maps find their ground-truths. The results in the second column of Table 4 show that there are only 0.09-0.37 percent of the data finding their ground-truths as pseudo labels in the training set. We then further evaluate the semantic parsing transformation performance. Table 4 illustrates there is no significant difference in semantic transformation performance with different N, though they are inferior to that trained with ground-truth semantic labels (the last row). However, it is not necessary for our model to generate semantic maps as accurate as possible, because our end-to-end training strategy is able to refine the predicted semantic maps, as analyzed in Fig. 9. In our experiments, we set N as 500.

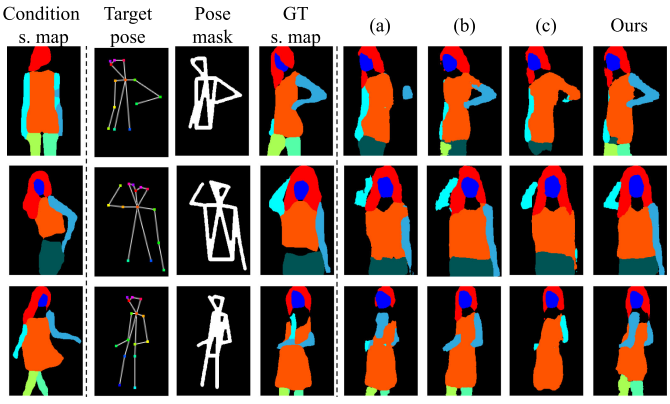


Fig. 10. Analysis of semantic parsing transformation. (a) Remove pose masks from the input. (b) Remove pose heat maps from the input. (c) Remove $\mathcal{L}_S^{adv}$ in Eq. (5). The results of the semantic maps generated by our semantic generative network are in the right.

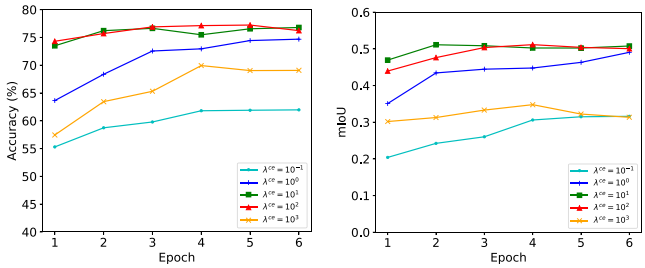
• *Loss Functions.* In semantic parsing transformation, we mainly explore the effectiveness of adversarial loss and its parameter setting.

Effectiveness of Adversarial Loss $\mathcal{L}_S^{adv}$. In the loss function (Eq. (5)) for semantic parsing transformation, $\mathcal{L}_S^{ce}$ is dispensable, because it provides straightforward supervision for the semantic generator to learn the transformation between different poses. But when trained only with $\mathcal{L}_S^{ce}$, the network would overlook the human body structure and generate unrealistic results, as shown in Fig. 10c. However, the introduction of $\mathcal{L}_S^{adv}$ helps generate high-quality and realistic semantic maps (see *Ours* in Fig. 10).

Parameter Setting . We further explore the influence of λ^{ce} in Eq. (5). To evaluate different λ^{ce}, we randomly select 1,000 samples, then calculate accuracy and mIoU with the predicted and ground truth semantic maps. The results are shown in Fig. 11. When λ^{ce} is too small (i.e., $\lambda^{ce} = 10^{-1}$), the semantic generative network is not powerful enough to learn the transformation between different poses. With a larger λ^{ce}, the accuracy and mIoU improve accordingly, and the network converges more quickly. However, when $\lambda^{ce} = 10^3$, the weaker constraints by $\mathcal{L}_S^{adv}$ lead to degradation in generating accurate semantic maps. Note that the results in Fig. 11 are just indicative to help us generate reasonable semantic maps more efficiently. On the one hand, ground truth semantic labels are not available to fine-tune the human parser [11], so there are errors in ground truth semantic maps, which could influence the results of

TABLE 4
Analysis of Pseudo Label Generation

N	Ground-Truth (%)	Accuracy (%)	mIoU
100	0.09	73.57	0.476
500	0.29	75.01	0.473
1000	0.37	74.56	0.474
Ground-Truth	–	76.66	0.495



(a) Accuracy (%) (b) mIoU

Fig. 11. Results of predicted semantic maps in terms of accuracy and mIoU with different λ^{ce}.

accuracy and mIoU. On the other hand, we do not have to get predicted semantic maps the same with the ground truths. In our experiments, we set λ^{ce} as 10^2.

4.3.2 Analysis of Appearance Generation

To analyze the appearance generation module, we show an ablation study for foreground generation and background generation, respectively.

• *Foreground Generation*. We first explore the effectiveness of different loss terms for foreground generation, and then analyze the effects of coarser semantic maps on generated foregrounds. Note that to avoid the influence of semantic map prediction, we conduct experiments with TS-GT model when evaluating $\mathcal{L}_A^{sty}$ and $\mathcal{L}_A^{face}$.

Effectiveness of Style Loss $\mathcal{L}_A^{sty}$. In Fig. 12a, the semantic-aware style loss $\mathcal{L}_A^{sty}$ is replaced with a mask-style loss, which refers to replace semantic maps in Eq. (10) with binary masks as Fig. 3. The rectangular masks estimated from body joints are not able to locate body parts accurately, leading to dizzy contour in the generated images. In Fig. 12b, the semantic-aware style loss $\mathcal{L}_A^{sty}$ is replaced with the patch-style loss defined by [1]. It only enforces the textures around corresponding pose joints similar, overlooking those of main body parts. Thanks to our semantic-aware style loss, our network is able to transfer the textures from corresponding semantic regions accurately.

Effectiveness of Face Loss $\mathcal{L}_A^{face}$. Without the constraints by the face discriminator, the network is difficult to handle facial structures and generate natural faces, which can be verified in Fig. 12c. However, our results with $\mathcal{L}_A^{face}$ show that face loss effectively helps generate realistic-looking faces, and further present more pleasing visual quality of output images.

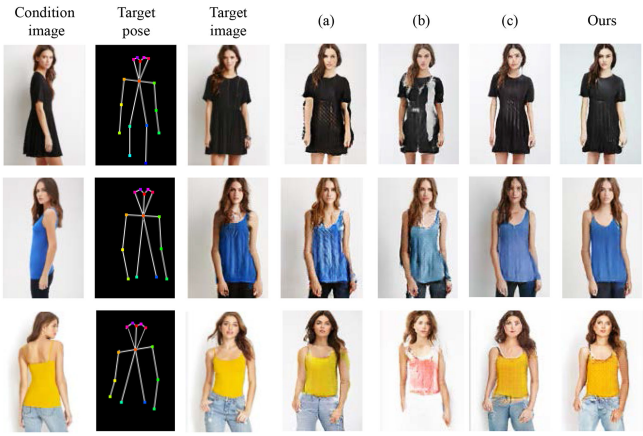


Fig. 12. Analysis of style loss $\mathcal{L}_A^{sty}$ and face loss $\mathcal{L}_A^{face}$ in appearance generation. (a) $\mathcal{L}_A^{sty}$ is replaced with mask-style loss. (b) $\mathcal{L}_A^{sty}$ is replaced with patch-style loss. (c) Without $\mathcal{L}_A^{face}$. The results of TS-GT with full losses are shown in the right.

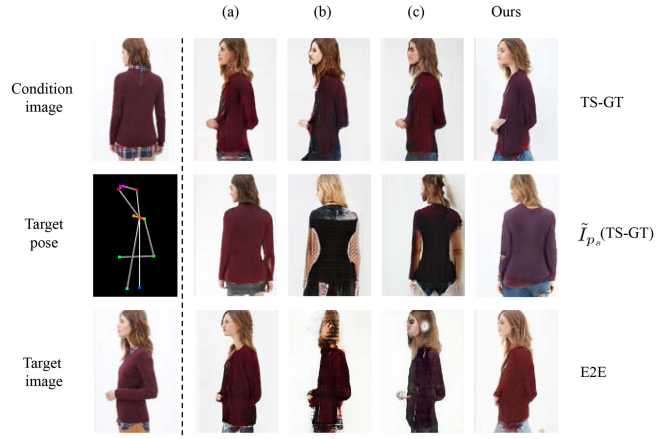


Fig. 13. Analysis of the losses on $\tilde{I}_{p_s}$ in appearance generation. (a) w/o $\mathcal{L}'_{sty}$, (b) w/o $\mathcal{L}'_{adv}$, (c) w/o $\mathcal{L}'_{sty}$ and $\mathcal{L}'_{adv}$. The results of our full losses are in the right.

Effectiveness of Losses on $\tilde{I}_{p_s}$. In our work, the loss penalizes style and adversary for the recovered image $\tilde{I}_{p_s}$, even though it could be directly compared with the condition image I_{p_s}. In Fig. 13, we give an ablation study to confirm the impact of losses on the recovered image $\tilde{I}_{p_s}$. We denote the style loss on the recovered image as $\mathcal{L}'_{sty}$ and adversarial loss as $\mathcal{L}'_{adv}$. The two losses do not have much impact on the final results in TS-GT (1st row). But the absence of adversarial loss leads to degradation on the recovered image (2nd row). It further results in unstable end-to-end training (3rd row).

Effects on Coarser Semantic Maps. We investigate the effects of semantic maps for appearance generation. We conduct experiments with semantic maps under different qualities, by downsampling with different scales. The results can be seen in Fig. 14. Our appearance generation module is able to handle minor errors in the semantic maps when downsampled 2 or 4 times. The coarser semantic maps with 8 or 16 times downsampling lead to unrealistic results, largely due to unrealistic shape prior in the semantic maps.

• *Background Generation*. We further analyze the effectiveness of our background generative network and compare different background generation schemes in Figs. 15 and 16, respectively.

Effectiveness of Our Background Generative Network. Without explicitly rendering background as our conference

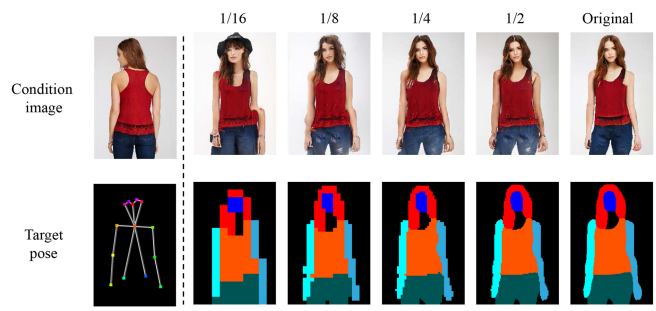


Fig. 14. Results with semantic maps in different qualities.

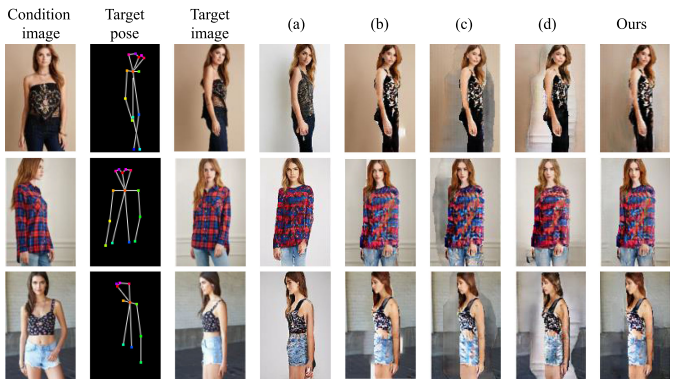


Fig. 15. Analysis of the proposed background generative network and our iterative training strategy. (a) W/o background generative network [12]. (b) Baseline results for background generation with PatchMatch [56] for background generation. (c) Training an independent inpainting network [57] using auxiliary data. (d) Training the proposed model without auxiliary data. Our final results are in the right.

version [12], the model generates background according to the distribution of training dataset, ignoring the background condition in the input image. The results can be seen in Fig. 15a. A straightforward strategy for background generation is to use an inpainting method. We use PatchMatch [56] as a background generation baseline, and the results are shown in Fig. 15b. We observed PatchMatch sometimes fails [56] (see the holes in Fig. 15b) because the large hole makes it hard to find an optimal matched patch in the visible region. We also train a background generative network [57] independently with human-shape masks on auxiliary data and the results are in Fig. 15c. There are always artifacts in the generated backgrounds. It is because the images in auxiliary data are cluttered, while the backgrounds in original images are clean and structured. The domain gap between auxiliary data and original images leads to degradation in the inpainted backgrounds. In this work, we train the auxiliary and original images jointly for background generation. The results are in the right of Fig. 15. We further conduct an experiment that jointly trains

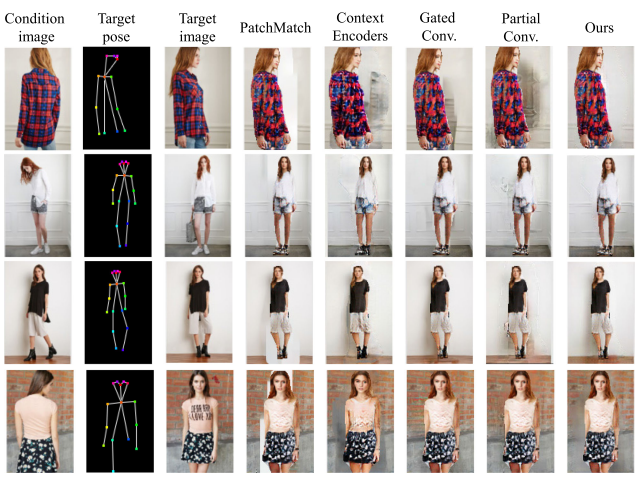


Fig. 16. Comparison with different background generation methods, including PatchMatch [56], context-encoders [57], gated-convolution based inpainting network [58], and partial-convolution based inpainting network [59]. Our results are in the right. We show a failure case in the last row which can be improved with a stronger backbone for the background generative network.

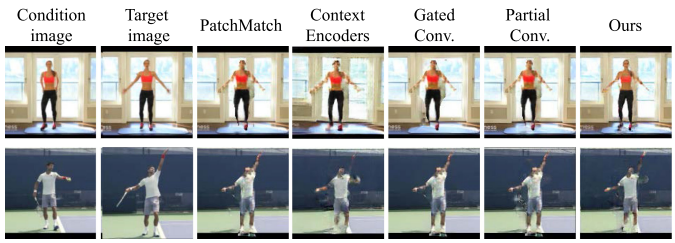


Fig. 17. In-the-wild results. We try images from the Posewarp dataset [7].

the background and foreground networks without auxiliary data. The results in Fig. 15d illustrate it is difficult for the network to predict missing regions with coherent textures.

Comparison With Different Background Generation Methods. In Fig. 16, we compare more results based on different inpainting methods, including PatchMatch [56], context-encoders [57], gated-convolution based inpainting network [58], and partial-convolution based inpainting network [59]. Results show that PatchMatch [56] fails to fill the holes as illustrated above. For the deep network based methods [57], [58], [59], the works in [58], [59] are better at generating coherent textures and seamlessly rendering the boundary for the inpainted regions compared to context-encoders [57]. However, we still observe similar degraded results in the backgrounds due to the domain gap between the auxiliary and original images (see the 1st-3rd rows in Fig. 16). Our training scheme, however, inpaints the background successfully and smoothly. We also show a failure case in Fig. 16 (the last row), that [58], [59] achieve more satisfactory results. We argue that our result can be further improved with a stronger backbone for the background generative network with our iterative training strategy.

In-the-Wild Results. To further evaluate our background generative network, we apply our model on more general images from the Posewarp dataset [7] with more complex backgrounds. The images are shown in Fig. 17. Compared with other state-of-the-art inpainting methods ([56], [57], [58], [59]), the results suggest the superiority of our model, not only generating realistic foreground but also smooth background.

4.4 Applications

The pose-guided generation results indicate that our model essentially learns the mapping between corresponding semantic regions. It inspires us to apply our model on some semantic-aware image generation tasks. To demonstrate the versatility of our model, we show some interesting applications in the following.

- *Clothing Texture Transfer.* Clothing textures can be transferred from the condition image to the target image with their semantic maps. Fig. 18 presents the bidirectional transfer results. Compared to image analogy [60] and neural doodle [61], our model not only preserves and transfers textures, but also automatically generates photo-realistic faces.
- *Controlled Image Manipulation.* We are able to manipulate image generation by editing the semantic maps into the desired layouts. In the top of Fig. 19, we edit the sleeve lengths, and in the bottom we change the dress to pants. Compared to image analogy [60] and neural doodle [61],

Fig. 18. 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 results with image analogy [60] and neural doodle [61].

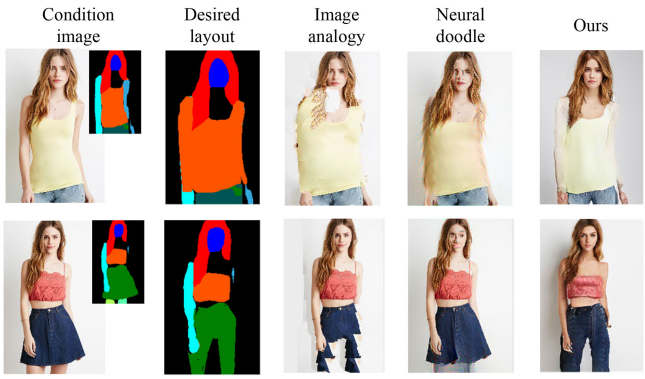
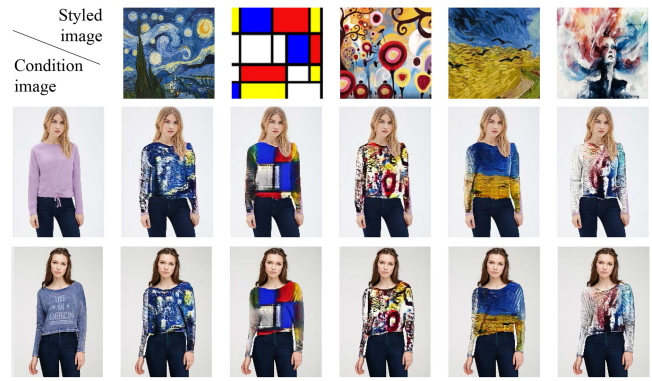


Fig. 19. Application for controlled image manipulation. By manually editing the semantic maps, we can generate images in the desired layout.

(a) Try-on with clothing.



(b) Try-on with styled images.

Fig. 21. Application for virtual try-on. Our model can synthesize clothing or styled images seamlessly on the condition images.

our model changes the appearance of the given image successfully according to the desired layout.

• *Virtual Try-On.* By feeding the product image into the appearance encoder and the semantic map of the clothed person into the semantic encoder, our appearance generation module is also applicable for the task of virtual try-on. In Fig. 21a, our model synthesizes photo-realistic new images and overlays the product image seamlessly onto the corresponding region of the clothed person. We also try to transfer the styled images on the clothing. The results can be seen in Fig. 21b.

Implementation Details. All the applications above are based on the foreground generative network in the appearance generation module, which is trained on the DeepFashion dataset [13] for our unpaired pose-guided image generation. No additional training on product images is required. For clothing texture transfer and controlled image manipulation, we feed the corresponding inputs to the foreground generator in Fig. 4. For virtual try-on, we need to define the inputs for clothing/styled images to adapt to the foreground generator. Taking try-on with clothing as an example, Fig. 20 shows our detailed implementation. We need to provide the coordinates of the clothing landmarks $\mathbf{p}_c$, including *left/right collar* , *left/right elbow* , *left/right sleeve end* , and *left/right hem* . Besides, the inputs for E_A^F include the clothing image I_c and its semantic map S_c. Note that the semantic label for the clothing image is the same as the try-on region in the condition image I_t. The inputs for E_S' include the pose map $\mathbf{p}_t$ and the semantic map S_t from I_t. If the clothing attributes of I_c and I_t are inconsistent (i.e., I_c has short sleeves while I_t has longer ones), we need to manipulate S_t to the desired layout as controlled image manipulation. Then we have $I'_{out} = G_A^F\big(E_A^F(\mathbf{p}_c, S_c, I_c), E_S'(\mathbf{p}_t, S_t)\big)$. In virtual try-on, it is usually important to keep the identity of the target person. In our implementation, we generate a binary mask M_f from S_t for the try-on region. To keep the original face, we have $I_{out} = I'_{out} \otimes M_f + I_t \otimes M_b$, where $M_b = \mathbf{1} - M_f$. Similarly, for try-on with styled images, we need to define the corners of the styled images as $\mathbf{p}_c$ and generate a semantic map as S_c, so the model can map the styled textures on the condition image correspondingly.

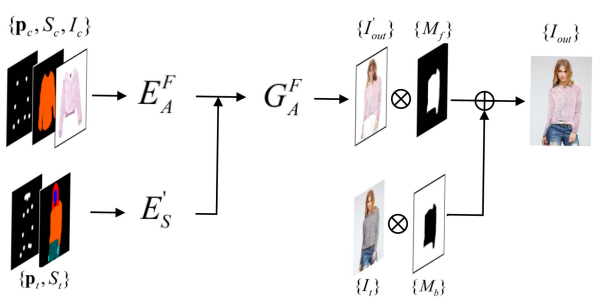


Fig. 20. Implementation details for try-on with clothing.

4.5 Failure Cases and Open Issue Discussion

Although our model generates impressive results, we also observe some failure cases as shown in Fig. 22. In the condition semantic map of the first example, the human parser incorrectly parses the sleeves as arms. Such error makes the semantic generative network unable to predict the semantic map properly. In the second example, the generated

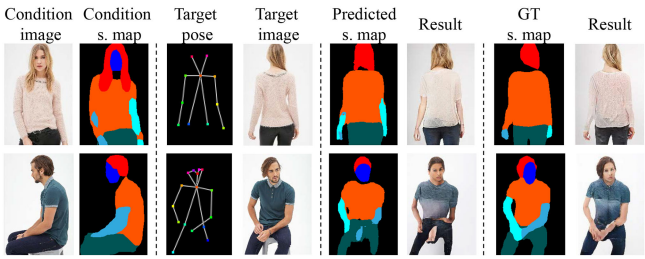


Fig. 22. Some failure cases by our model.

semantic map is less satisfactory because the transformation to a rare pose is very complex, leading to less realistic generated images. However, with ground truth semantic maps, we still obtain pleasing results from our model. These failure cases can be possibly solved by user interaction.

Based on the above failure cases and previous related works, we further discuss some open issues to inspire future work in this problem. First, though human semantic parsing provides crucial guidance for person image generation, the model can be vulnerable to parsing errors. Thus, a crucial component is improving robustness to semantic parsing, for instance, by jointly training the human parser and person image generation model. Second, the model does not work well with large pose changes, such as rare poses, or poses at different scales. Data augmentation could be an easy solution. It can also be regarded as a domain adaptation problem and new models are desirable. In the end, retaining background is a step toward generating temporally-smooth videos. Incorporating more spatial and temporal context may be important for pose-guided video generation.

5 Conclusion

In this paper, we propose a model for unpaired person image generation. To handle the complexity of learning a direct mapping between different poses, the hard problem is decomposed into semantic parsing transformation and appearance generation. A semantic generative network first predicts the semantic map of the desired pose explicitly. Then the appearance generation module respectively synthesizes the foreground and background, in which the foreground generative network renders semantic-aware textures, while the background generative network aims to retain the background from the condition image. To overcome the absence of ground truth images, we propose an iterative optimization procedure to train the appearance module, so that the background generative network can predict missing regions caused by pose changes. Besides, end-to-end training of the overall pipeline enables better prediction for semantic maps and further final results. Our model is also versatile on some interesting image generation applications, including clothing texture transfer, controlled image manipulation, and virtual try-on. However, our model may fail when there are errors in the condition semantic map. Finally, we discuss some open issues and possible future work in this problem.

Acknowledgments

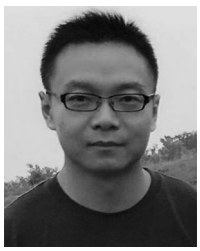
This work was supported by the National Natural Science Foundation of China under contract No. 61772043 and Beijing Natural Science Foundation under contract No. 4192025.

References

[1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 406–416.
[2] I. Goodfellow *et al.* "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[3] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8620–8628.
[4] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8857–8866.
[5] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3408–3416.
[6] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses for pose-based human image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 118–126.
[7] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8340–8348.
[8] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 99–108.
[9] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 955–10 964.
[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[11] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 932–940.
[12] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2357–2366.
[13] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1096–1104.
[14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.
[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.
[16] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Machine Learn.*, 2014, pp. 1278–1286.
[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.
[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.
[20] W. Xian, P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8456–8465.
[21] H. Huang *et al.*, "IntroVAE: Introspective variational autoencoders for photographic image synthesis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 52–63.
[22] I. Gulrajani *et al.*, "PixelVAE: A latent variable model for natural images," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.
[23] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 658–666.
[24] X. Chen *et al.*, "Variational lossy autoencoder," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.
[25] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–35.

[26] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.
[27] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium: GANs do not need to decrease a divergence at every step," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–21.
[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[29] Z. Yi, H. R. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2849–2857.
[30] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865.
[31] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7543–7552.
[32] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 135–12 144.
[33] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 217–225.
[34] S. Hong, X. Yan, T. S. Huang, and H. Lee, "Learning hierarchical semantic image manipulation through structured representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2708–2718.
[35] D. Lee, S. Liu, J. Gu, M.-Y. Liu, M.-H. Yang, and J. Kautz, "Context-aware synthesis and placement of object instances," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 10 393–10 403.
[36] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7986–7994.
[37] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1219–1228.
[38] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 474–484.
[39] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays, "SwapNet: Garment transfer in single view images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 666–682.
[40] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model of people in clothing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 853–862.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[42] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own Prada: Fashion synthesis with structural coherence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1680–1688.
[43] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–21.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[45] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.
[46] M. Jaderberg *et al.*,"Spatial transformer networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[47] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[48] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5505–5514.
[49] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[51] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[54] J. Johnson, A. Alahi, and L. Fei-Fei , "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[55] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[56] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 24.
[57] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.
[58] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4471–4480.
[59] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–100.
[60] A. Hertzmann, "Image analogies," in *Proc 28th Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 327–340.
[61] A. J. Champandard, "Semantic style transfer and turning two-bit doodles into fine artworks," 2016, *arXiv:1603.01768*.

Sijie Song (Student Member, IEEE) received the BS degree in computer science from Peking University, Beijing, China, in 2016, where she is currently working toward the PhD degree with the Wangxuan Institute of Computer Technology. Her research interests include computer vision and image processing.



Wei Zhang received the PhD degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, China, in 2015. He is currently a senior researcher with JD AI Research, Beijing, China. He was a visiting scholar with DVMM Group, Columbia University, New York, in 2014. His research interests include computer vision and multimedia, especially the visual content recognition and generation. He has won two competitions in FGVC 2019, the runner-up in TRECVID Instance Search in 2012, the Best Demo Award in ACM-HK openday 2013.

Jiaying Liu (Senior Member, IEEE) received the PhD (hons.) degree in computer science from Peking University, Beijing, China, in 2010. She is currently an associate professor with the Wangxuan Institute of Computer Technology, Peking University. She has authored more than 100 technical articles in refereed journals and proceedings, and holds 42 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a senior member of the CSIG and CCF. She was a visiting scholar with the University of Southern California, Los Angeles, from 2007 to 2008. She was a visiting researcher with the Microsoft Research Asia in 2015 supported by the Star Track Young Faculties Award. She has served as a member of Membership Services Committee in IEEE Signal Processing Society, a member of Multimedia Systems & Applications Technical Committee (MSA TC), Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society, a member of the Image, Video, and Multimedia (IVM) Technical Committee in APSIPA. She has also served as the associate editor of the *IEEE Transactions on Image Processing*, and the *Elsevier Journal of Visual Communication and Image Representation*, the technical program chair of IEEE VCIP-2019/ACM ICMR-2021, the publicity chair of IEEE ICME-2020/ICIP-2019, and the area chair of CVPR-2021/ECCV-2020/ICCV-2019. She was the APSIPA distinguished lecturer (2016-2017).

Zongming Guo (Member, IEEE) received the BS degree in mathematics, and the MS and PhD degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively. He is currently a professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include video coding, processing, and communication. He is the executive member of the China-Society of Motion Picture and Television Engineers. He was a recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013.

Tao Mei (Fellow, IEEE) received the BE and PhD degrees in electrical and computer engineering from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a technical vice president with JD.com and the deputy managing director of JD AI Research, where he also serves as the director of Computer Vision and Multimedia Lab. Prior to joining JD.com in 2018, he was a senior research manager with Microsoft Research Asia in Beijing, China, where he has shipped more than 20 inventions and technologies to Microsoft products and services. He has authored or coauthored more than 200 publications (with 12 best paper awards) in journals and conferences, ten book chapters, and edited six books. He holds more than 50 US and international patents. He was the recipient of a number of awards from prestigious multimedia journals and conferences, including the best paper awards from the *IEEE Transactions on Multimedia* (2019 and 2013), the *ACM Transactions on Multimedia Computing, Communications, and Applications* (2017), the *IEEE Transactions on Circuits and Systems for Video Technology* (2014), ACM International Conference on Multimedia (2009 and 2007), etc. He is or has been an editorial board member of the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *IEEE Transactions on Multimedia*, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, the *Pattern Recognition*, etc. He is the general co-chair of IEEE ICME 2019, the program co-chair of ACM Multimedia 2018, IEEE ICME 2015 and IEEE MMSP 2015. He is a guest professor of the University of Science and Technology of China, Fudan University, The Chinese University of Hong Kong (Shenzhen), and Yonsei University. He was elected as a fellow of IAPR in 2016, a distinguished scientist of ACM in 2016, and a distinguished industry speaker of the IEEE Signal Processing Society in 2017, for his contributions to large-scale multimedia analysis and applications.