Text Guided Person Image Synthesis

Xingran Zhou[1] Siyu Huang[1*] Bin Li[1] Yingming Li[1] Jiachen Li[2] Zhongfei Zhang[1]

[1] Zhejiang University [2] Nanjing University

{xingranzh, siyuhuang, bin_li, yingming, zhongfei}@zju.edu.cn, jiachen_li_nju@163.com

Figure 1: **Samples of text guided person image synthesis.** Given the reference images and the natural language descriptions, our algorithm correspondingly generates pose and attribute transferred person images. As shown in the left, our algorithm transfers the person pose based on *'He is walking forward to the camera'*, and also synthesizes shirts of various different colors. Similarly for the right example.

Abstract

This paper presents a novel method to manipulate the visual appearance (pose and attribute) of a person image according to natural language descriptions. Our method can be boiled down to two stages: 1) text guided pose generation and 2) visual appearance transferred image synthesis. In the first stage, our method infers a reasonable target human pose based on the text. In the second stage, our method synthesizes a realistic and appearance transferred person image according to the text in conjunction with the target pose. Our method extracts sufficient information from the text and establishes a mapping between the image space and the language space, making generating and editing images corresponding to the description possible. We conduct extensive experiments to reveal the effectiveness of our method, as well as using the VQA Perceptual Score as a metric for evaluating the method. It shows for the first time that we can automatically edit the person image from the natural language descriptions.

1. Introduction

Person images are produced in any time today due to the popularization of visual capturing devices such as mobile phones, wearable cameras, and surveillance systems. The demand for a user-friendly algorithm to manipulate person images is growing rapidly. In practice, people usually represent the concept about a person's appearance and status in a very flexible form, *i.e.*, the natural languages. It is our understanding that the guidance through text description for generating and editing images is a friendly and convenient way for person image synthesis.

In this paper, we propose a new task of *editing a person image according to natural language descriptions*. Two examples of this task are shown in Fig. 1. Given an image of a person, the goal is to transfer the visual appearance of the person under the guidance of text description, while keeping the invariance of person identity. Specifically, the pose, attributes (*e.g.*, cloth color), and the other properties of the identity are simultaneously edited to satisfy the description.

Generative Adversarial Networks (GANs) have offered a solution to conditioned realistic image generation. The text-to-image approaches [28, 22, 32, 30, 29] synthesize images with given texts without the reference images, where the semantic features extracted from texts are converted to the visual representations to constitute the generated images. However, most of these approaches only succeed in flower or bird image generation. In regard to person editing, the pose guided generation methods [18, 17, 5, 1] transfer the person pose by taking the target pose as input to instruct the generation process, while the editing of person image under the guidance of natural language descriptions has rarely

* Corresponding author

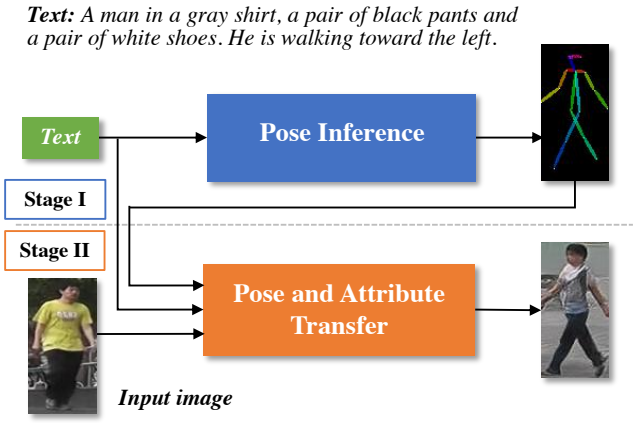

Figure 2: **Simple illustration of our approach.** In Stage-I, we infer a reasonable human pose from the natural language description. In Stage-II, our method takes the predicted pose, the reference image, and the text as input to synthesize a pose and attribute transferred person image by keeping the person identity.

been studied in the existing literature.

Motivated by these considerations, we propose a novel text guided person image synthesis framework, which is able to semantically edit the pose and the attributes of the person in consistence with the text description while retaining the identity of the person. As shown in Fig. 2, our method is comprised of two stage successively. The two stages are both built upon the adversarial learning conditioned on the text description. Specifically, Stage-I is a newly proposed pose inference network, in which a reasonable target pose is inferred from the text description as an intermediate product to lay the foundation for the subsequent pose transferring. A set of basic poses is drawn from the training dataset. The pose inference network first selects a basic pose with respect to the exact direction, and then refines every joint in details to conform to the text description. By the pose inference network, the target pose is guaranteed to model the shape and the layout of unique body pose structure of a person.

Stage-II takes the predicted pose, the reference image, and the text description as input to yield a realistic-look pedestrian image by manipulating both the pose and the appearance of the reference image. A multi-modal learning loss involved an attention mechanism is proposed to establish the link among different words in the text and the subregions in the image. Moreover, a novel attention upsampling module is developed in this stage for better combining the pose feature and the semantic embedding. Compared with the previous image editing methods, our model is able to simultaneously manipulate multiple person attributes, enabling a more interactive and flexible approach to person image synthesis.

The contribution of this paper is listed as follows. 1) We propose a new task of manipulating person images based on natural language descriptions, towards the goal of user-friendly image editing. 2) For the first time, we propose a GAN-based pose inference network to generate the human pose according to the text description, to the best of our knowledge. 3) We present a novel two-stage framework for text guided person image synthesis, in which the modules of attention upsampling and multi-modal loss are introduced to establish semantic relationships among images, poses, and natural language descriptions. 4) We propose the VQA Perceptual Score to evaluate the correctness of attribute changes corresponding to specific body parts.

2. Related works

Deep generative models. In recent years, deep generative models including Generative Adversarial Networks (GANs) [6], Variational Auto-encoders (VAEs) [12], and Autoregressive (AR) models [25] have attracted wide interests in the literature. The advances of generative models also drive further studies on generating and translating images, including the image to image translation [8, 34, 2], super-resolution [13, 20, 10], and the style transfer [31, 35, 9]. These techniques are of great importance for the computer vision research and with a plenty of applications.

Person image generation. Recent work has achieved impressive results in generating person images in the expected poses. For instance, Ma *et al*. [18] propose Pose Guided Person Image Generation (PG^2) which initially generates a coarse image and then refines the blurry result in an adversarial way. Balakrishnan *et al*. [1] present a modular generative network which separates a scene into various layers and moves the body parts to the desired pose. Ma *et al*. [17] use a disentangled representation of the image factors (foreground, background, and pose) to composite a novel person image. Esser *et al*. [5] present a conditional U-Net shape-guided image generator based on VAE for person image generation and transfer. It is desirable to edit and manipulate person images according to natural language descriptions.

Text conditioned generation. Reed *et al*. [22] first propose an end-to-end architecture based on conditional GANs framework, which generates realistic 64×64 images for birds and flowers from natural language descriptions. Their follow-up work [23] is able to generate 128×128 images by incorporating additional annotations of object parts. The StackGAN [30, 29] is proposed to generate natural images by utilizing a stacked structure consisting of multiple generators and discriminators to generate images of different sizes. Tao *et al*. [28] employ the attention mechanism into this problem into their solution, which is able to synthesize images with fine-grained details from the text. Another line of literature concentrates on editing images by natural language description. For instance, Dong *et al*. [4] manipulate

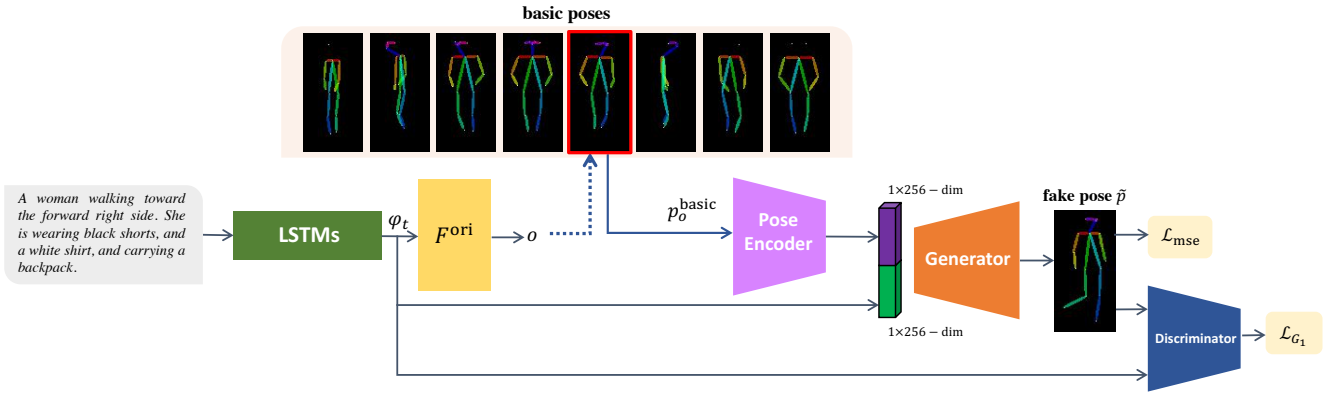

Figure 3: **Stage-I: Text guided pose generator.** We conclude the prior about poses in the training dataset as a series of basic poses. We first predict the orientation of the pose from the text by the orientation selection net F^o. Then, we train a single generator $G1$ that learns to manipulate every joint in the pose for fine-grained details.

images semantically with text descriptions. Nam *et al*. [19] enhance fine-grained details by learning disentangled visual attributes from text-adaptive discriminator. However, most of them only succeed in the flower or bird image generation. In this paper, we present a text guided person image synthesis framework which can generate and edit the person pose and attribute according to natural language text while retaining the identity of the person.

3. Method

3.1. Problem Definition

Our goal is to simultaneously transfer the pose and the appearance of a person in the reference image corresponding to the given text description.

Training data. For each person in the training dataset, there is a tuple (x, x', p, t) containing the source (reference) image x and the target image x' of the same identity with a different pose. p and t are the pose and the text description of x', respectively.

Our pipeline. To tackle this challenging problem, we factorize it into two stages:

- **Stage-I:** We infer a reasonable pose based on the given text t. (See Sec. 3.2)
- **Stage-II:** We generate a person image in which the pose and the attribute details of that person are changed according to target pose p and text t. (See Sec. 3.3)

3.2. Text Guided Pose Generator

In Stage-I, we propose a novel approach (see Fig. 3), named *text guided pose generator*, to infer a reasonable pedestrian pose satisfying the description.

We obtain the prior about poses in the training dataset as the basic poses and manipulate joints in these poses. Generally, the direction of a target pose is first estimated based on the description, then the target pose is generated in conjunction with detailed fine-tuning.

Basic poses. Synthesizing a pose directly from the text is difficult, as both the orientation and the other details (*e.g*., motions, posture) of a pose need to be considered. Following [21], we group the poses of all training images into K clusters and compute the mean pose p_i^{basic} of the i-th cluster, forming a basic pose set $\{p_i^{\text{basic}}\}_{i=1}^{K}$ (see Fig. 3 for basic poses, where we use $K = 8$ like [21]). We assume that the basic poses orient toward all K different directions.

Pose inference. Given the text description t corresponding to the target image x', we take the output of the final hidden layer of LSTMs as the sentence representation vector φ_t. We predict the orientation of the pose, $o = \arg\max_o F^{\text{ori}}(\varphi_t), o \in \{1, ..., K\}$. F^{ori} is the orientation selection net implemented as fully-connected layers. The basic pose p_o^{basic} which matches the orientation o is selected from the K basic poses.

We observe that verbs in the text can be vague in specifying the specific action of limbs. For example, the word *walking* does not specify which leg to stride. The predicted pose by a regression method could either be striding on both legs or staying upright. Therefore, we train a single generator G_1 that learns to adjust the details of the pose, formulated as $G_1(p_o^{\text{basic}}, \varphi_t) \rightarrow \tilde{p}$. The discriminator D_1 outputs a probability that a pose is real conditioned on the text. D_1 forces G_1 to concern about posture details depicted by the text consistent with the real pose. The adversarial loss of discriminator D_1 is defined as

$$\begin{aligned}\mathcal{L}_{D_1} = & -\mathbb{E}_{p\sim \Pr(p)}[\log D_1(p)] \\ & -\mathbb{E}_{\tilde{p}\sim \Pr(\tilde{p})}[\log(1 - D_1(\tilde{p}))]\end{aligned} \tag{1}$$

And the adversarial loss of generator G_1 is

$$\mathcal{L}_{G_1} = -\mathbb{E}_{t\sim p_{data}}[\log D_1(G_1(p_o^{\text{basic}}, \varphi_t))] \tag{2}$$

However, we find that only using the adversarial loss makes the generated poses lack of pedestrian pose structure, as the values of pose heat-map are 1 merely within the

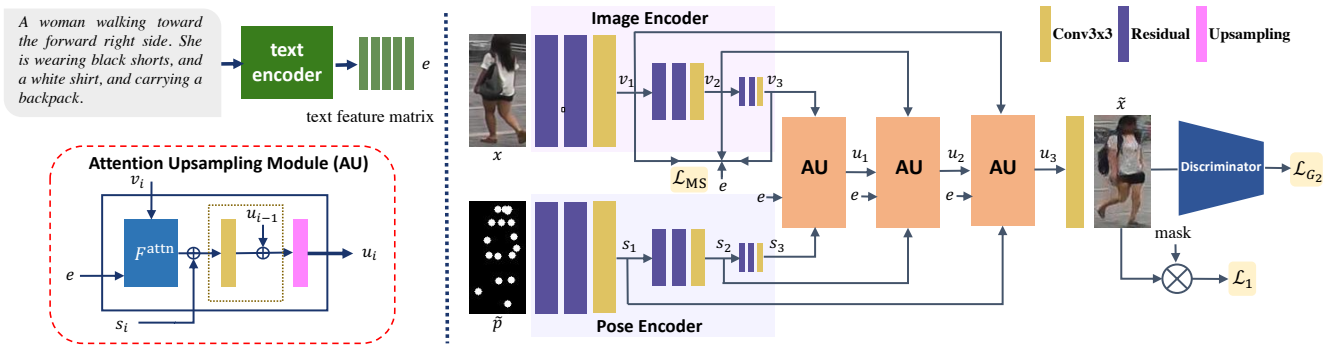

Figure 4: **Stage-II: Pose and attribute transferred person image generator.** It is a multi-modal learning scheme which builds the link between modalities of image, text, and pose. In addition, we propose a basic attentional upsampling (AU) module to better incorporate information of different modalities and spatial scales into image generation. The conjunction symbol in the AU module means the concatenation operation.

radius while the rest are almost 0. Thus, we add a mean-square-error item, $\mathcal{L}_{\text{mse}} = \|\tilde{p} - p\|^2$, to the adversarial loss of generator G_1, helping maintaining the unique structure.

The objective function of text guided pose generator is finally formulated as

$$\mathcal{L}_{\text{Stage-I}} = \mathcal{L}_{G_1} + \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 \mathcal{L}_{\text{cls}} \tag{3}$$

Here, λ_1 and λ_2 are hyper-parameters for balancing the three terms of Eq. (3). $\mathcal{L}_{\text{cls}}$ is the cross entropy between the estimated orientation o and the real orientation o_{real}.

3.3. Pose and Attribute Transferred Person Image Generator

We have predicted the target pose $\tilde{p}$ based on the text t so far. In Stage-II, our goal is to transfer the pose to the target pose $\tilde{p}$ and edit the appearance (*e.g.*, cloth color) according to certain key words in description t. [1] To tackle this challenging problem, we propose a multi-task *pose and attribute transferred image generator*, which is shown in Fig. 4.

Our multi-task person image synthesis framework is built upon the encoder-decoder structure.

- The *image encoder* extracts the image feature maps $(v_1, v_2, ..., v_m)$ of different scales by taking the source image x as input. $v_i \in \mathbb{R}^{l_i \times h_i \times w_i}$, where l_i, h_i, w_i are the dimension, height, and width of the feature map at the i-th scale, $i \in [1, ...m]$, m is the total number of downsampling in the encoder.
- The *text encoder* is a bi-directional LSTM which extracts the text feature matrix $e \in \mathbb{R}^{L \times N}$ of text t. L is the dimension of hidden states, N is the number of words. e is concatenated by the hidden states $(h_1, h_2, ..., h_N)$ corresponding to every word in t.
- The *pose encoder* [17] extracts the pose feature representations $(s_1, s_2, ..., s_m)$ of different scales by taking the target pose $\tilde{p}$ as input, similar to the image encoder, $s_i \in \mathbb{R}^{l_i \times h_i \times w_i}$.

Text-to-visual attention. We take the text feature matrix e and the image feature map v_i as input to calculate a dynamic text-to-visual attention which indicates the trend that each word takes care of each local visual region when generating images. The text-to-visual attention at the i-th scale is calculated as

$$F_i^{\text{attn}}(\hat{e}_i, \bar{v}_i) = \hat{e}_i \text{Softmax}(\hat{e}_i^\top \bar{v}_i) \tag{4}$$

where the visual feature $v_i \in \mathbb{R}^{l_i \times h_i \times w_i}$ is reshaped to $\bar{v}_i \in \mathbb{R}^{l_i \times h_i w_i}$, and the text feature matrix e is converted to a common semantic space $\hat{e}_i$ by an embedding layer as $\hat{e}_i = W_i e, W_i \in \mathbb{R}^{l_i \times L}$.

Attentional upsampling. We propose a basic module, named attentional upsampling (AU). The motivation is that our pose and attribute transfer problem contains multiple modalities of data (image, pose, and text). We apply this module to better incorporate the text-to-visual attention features and the pose features at different scales. The pose features conduct the layout and the structure, while the text-to-visual attentional features integrate attribute information from words into visual representation. In our experiments, we observe that the module is capable of transferring the pose and the attribute appearance of a person in the source image, while keeping the invariant identity of the source image and the generated image.

Our attentional upsampling operates on the image feature maps and pose feature maps at the same scale (see Fig. 4 Attentional Upsampling). To better retain information in the source image, the generators for image synthesizing and upsampling are weight-sharing, in which the fused features of different scales correspond to lower resolution

[1] The attribute that we focus on mainly is the color of clothes in this work, while in principle, our method can be easily extended to accommodating other attributes.

to higher resolution, respectively. The total m attentive operations in the upsampling correspond to those in the downsampling.

By using Eq. (4), we calculate the text-to-visual attention at the i-th scale as $z_i = F_i^{\text{attn}}(\hat{e}, \bar{v}_i)$. Then, z_i, s_i and the previous upsampling result u_{i-1} are incorporated and upsampled by

$$u_i = F_i^{\text{up}}(z_i, s_i, u_{i-1}) \tag{5}$$

For the smallest scale (*i.e.*, $i = 1$), z_1 and the pose feature s_1 are concatenated and upsampled as $u_1 = F_1^{\text{up}}(z_1, s_1)$. In such a recursive manner, the information of all different scales is incorporated in the final attentional upsampling result u_m. u_m is passed through a ConvNet to output the generated image $\tilde{x}$. In practice, we implement F^{up} as ConvNets with a nearest neighbor upsampling.

Multimodal loss. The multimodal loss function helps to establish the mapping between every word in the text and regions of images at different scales. The multimodal loss impose alignment among them for subsequently transferring appearance controlled by the text.

Similar to Eq. (4), the visual-to-text attention is calculated by

$$c_i = \hat{v}_i \text{Softmax}(\hat{v}_i{}^\top e) \tag{6}$$

The visual feature v_i is first reshaped to $\bar{v}_i \in \mathbb{R}^{l_i \times h_i w_i}$ and then converted to a common semantic space as $\hat{v}_i = U_i \bar{v}_i, U_i \in \mathbb{R}^{L \times l_i}$. $c_i \in \mathbb{R}^{L \times N}$, where the j-th column of c_i denotes the visual text attention for the j-th word at the i-th scale.

Inspired by [28], we calculate the similarity between the visual-to-text representation and the text feature matrix. The multi-scale visual-to-text distance is

$$\mathcal{D}(Q, T) = \sum_{i=1}^{m} \log\Big(\sum_{j=1}^{N} \exp(r(c_{ij}, e_j))\Big) \tag{7}$$

where Q refers to the image (query) and T refers to the description. $r(\cdot, \cdot)$ is the cosine similarity between two vectors, m is the number of scales.

For a batch of our training pairs $\{(x_i, t_i)\}_{i=1}^{I}$, we calculate the multi-scale visual-to-text distance matrix Λ; the element $\Lambda_{(i,j)} = \mathcal{D}(x_i, t_j)$. Following [28], the posterior probability that the text t_i matches with the image x_i is calculated as $P(t_i|x_i) = \text{Softmax}(\Lambda)_{(i,i)}$. Similarly, the posterior that the image x_i matches the text t_i is $P(x_i|t_i) = \text{Softmax}(\Lambda^\top)_{(i,i)}$.

The multimodal similarity $\mathcal{L}_{\text{MS}}$ measures the interaction responses for the pairing of sentences and images in a batch

$$\mathcal{L}_{\text{MS}} = -\sum_{i=1}^{I} \log P(t_i|x_i) - \sum_{i=1}^{I} \log P(x_i|t_i) \tag{8}$$

Multi-task person image generator. The total objective function of our multi-task person image generator is defined as

$$\mathcal{L}_{\text{Stage-II}} = \mathcal{L}_{G_2} + \gamma_1 \mathcal{L}_1 + \gamma_2 \mathcal{L}_{\text{MS}} \tag{9}$$

where γ_1 and γ_2 are hyper-parameters. $\mathcal{L}_1$ is the L1 distance between generated image $\tilde{x}$ and real image x', written as

$$\mathcal{L}_1 = \|(\tilde{x} - x') \odot M\|_1 \tag{10}$$

where M is the mask of the target pose [18]. We use three conditional probabilities to improve the quality of the generated images. The adversarial loss for the generator G_2 is defined as

$$\mathcal{L}_{G_2} = \underbrace{-\mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log D_2(\tilde{x}, e)]}_{\text{text conditional loss}} \underbrace{-\mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log D_2(\tilde{x}, p)]}_{\text{pose conditional loss}} + \underbrace{-\mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log D_2(\tilde{x}, e, p)]}_{\text{text and pose conditional loss}} \tag{11}$$

and the adversarial loss for the discriminator D_2 is

$$\begin{aligned}\mathcal{L}_{D_2} = &\underbrace{-\mathbb{E}_{x'\sim p_{data}}[\log D_2(x', e)] - \mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log(1 - D_2(\tilde{x}, e)]}_{\text{text conditional loss}} + \\ &\underbrace{-\mathbb{E}_{x'\sim p_{data}}[\log D_2(x', p)] - \mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log(1 - D_2(\tilde{x}, p)]}_{\text{pose conditional loss}} + \\ &\underbrace{-\mathbb{E}_{x'\sim p_{data}}[\log D_2(x', e, p)] - \mathbb{E}_{\tilde{x}\sim \Pr(\tilde{x})}[\log(1 - D_2(\tilde{x}, e, p)]}_{\text{text and pose conditional loss}}\end{aligned} \tag{12}$$

3.4. VQA Perceptual Score

The evaluation metrics of GANs in the existing literature are not specifically designed for the attribute transfer task.

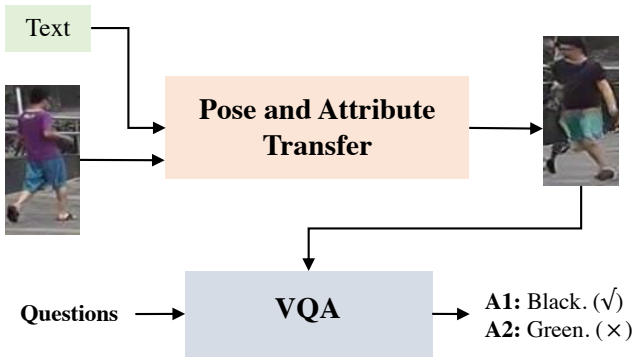

Figure 5: **Illustration of VQA perceptual Score.** The accuracy of answers returned by the VQA model denotes the attribute transfer correctness of generative models.

The Inception Score (IS) [24] measures the authenticity of synthesis and Structural Similarity (SSIM) [26] measures

the structural integrity of images. To this end, we propose a novel metric, named VQA perceptual score, for the assessment of the attribute transfer correctness, *i.e.*, whether the attributes of a person in the generated images are in agreement with the text description.

We first generate images using the method we propose by randomly changing the color adjectives of the clothes in the text (10 colors are considered). Correspondingly, the color word is recorded as the correct answer. Then a related question is automatically generated about the body part (shirt, pants, *etc.*) and its color. We ask the VQA model [11] with the question and the image. Finally, we gather the responses from the VQA model and calculate the accuracy, *i.e.*, the VQA perceptual score. Assuming that T is the number of images which receive all the correct answers from the VQA model, and that there are N images in total, the VQA perceptual score is defined as $\frac{T}{N}$.

4. Experiments

4.1. Dataset

CUHK-PEDES dataset [14] is the only caption-annotated pedestrian image dataset as far as we know. The dataset contains 40,206 images of 13,003 persons collected from five person re-identification datasets, CUHK03 [16], Market-1501 [33], SSM [27], VIPER [7], and CUHK01 [15]. Each image in the dataset is annotated with descriptions by crowd-sourcing.

In order to train the text guided pose generator, we add some phrases which describe the orientation, since the original descriptions rarely contain the information. An orientation phrase is an important guidance because otherwise the orientation in the generated image can be arbitrary and randomly different when lacking the orientation information. This may bring troubles to both the model training and testing.

For each image, a short phrase is appended according to the result of the clustering mentioned in Sec. 3.2. Every phrase corresponds to one of the $K = 8$ basic orientations. We have manually checked the phrases to ensure a high quality dataset.

Following [18], the identities of training set and testing set are exclusive. All images in the dataset are resized to 128×64. In the training set, we have 149,049 pairs each of which is composed of two images of the same identity but different poses. We have 63,878 pairs in the testing set.

4.2. Comparison with Baselines

As there is no existing work exactly comparable with this work, we implement four different baselines with appropriate modifications to make them comparable with our model as follows.[2]

[2]We do not use any extra pre-trained models in our framework, such that all the parameters in our model are trained from scratch, which is different from [28].

1. Modified Semantic Image Synthesis (SIS) [4] (mSIS). SIS uses a plain text encoder without our proposed attentional upsampling module. SIS only edits the attribute, but the pose is not involved. We append a pose encoder to it for pose transfer. The generator synthesizes a new person image based on the encoded reference image feature and the conditional representations of the target pose and the target text description.

2. Modified AttnGAN [28] (mAttnGAN). We add an image encoder and a pose encoder to the original AttnGAN [28, 19]. Specifically, an additional `inception_v3` network is adopted to establish the link among different words in the text and the subregions in the image.

3. Modified PG^2 [18] (mPG^2). Pose guided person image generation (PG^2) only generates the pose transferred person image. In this baseline, we append a text encoder for attribute transfer. Our multi-task problem is separated into two single-task problems, in which the pose transferred image is first synthesized and then the image is edited according to the text description step by step.

4. Single attentional upsampling (SAU). It conducts only $m = 1$ attentional upsampling module at the smallest scale, serving as an ablation study for our complete attentional upsampling modules.

Quantitative analysis. Following [17], we use the Inception Score (IS) [24] and the Structural Similarity (SSIM) [26] to measure the quality of generated person images. We evaluate the IS on tasks of pose transfer (PT) and pose and attribute transfer (P&AT). We only evaluate the SSIM on PT, as the SSIM is calculated based on images' mean and variance values which vary during attribute transferring.

We evaluate the four baselines and our method on metrics of IS and SSIM, as shown in Table 1. We can see that mSIS, mAttnGAN, and mPG^2 are the improved variants of the existing methods while their IS and SSIM values are lower than that of our model. It indicates that simply replenishing the rest procedure of the existing methods may not be feasible for the challenging problem proposed in this paper. SAU is better than the other baselines, while it is also worse than our complete framework. It indicates that the attentional upsampling module proposed in this work enables a robust learning of pose and attribute transferring.

VQA perceptual score. The VQA perceptual score of our model and the baselines are shown in Table 2. mSIS gains a relatively high score, while its generated images almost lose the human structure which is intolerable for visual effects. The scores of mAttnGAN and mPG^2 are relatively low, confirming that a separate training of the two tasks is

Figure 6: Four examples of text guided person image synthesis by our model. The first two columns are reference and target images (*i.e.*, ground truth (GT)) from the training set. The third column is the target image generated by our model. We additionaly show that our model is capable of transferring attribute if we dedicate to changing the attribute (*e.g.*, color) in the description.

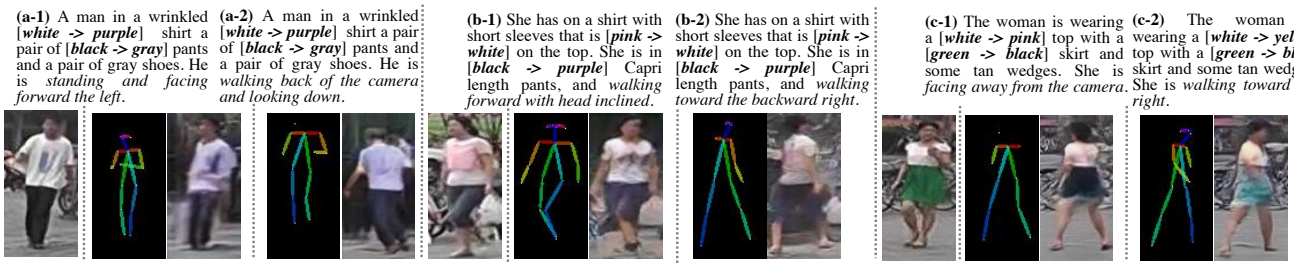

Figure 7: Interactive editing. By arbitrarily changing input words, our model can change a person to different poses for the same reference image. Our model can also transfer different attributes in one image at the same time, indicating a flexible and robust image generation procedure.

Model	SSIM (PT)	IS (PT)	IS (P&AT)
mSIS	0.239 ± .106	3.707 ± .185	3.790 ± .182
mAttnGAN	0.298 ± .126	3.695 ± .110	3.726 ± .123
mPG^2	0.273 ± .120	3.473 ± .009	3.486 ± .125
SAU	0.305 ± .121	4.015 ± .009	4.071 ± .149
Ours	**0.364** ± .123	**4.209** ± .165	**4.218** ± .195

Table 1: The SSIM score for pose transfer, and the IS for pose transfer and pose & attribute transfer (higher is better).

hard to achieve a balance between transfers of pose and attribute. Our model jointly addresses the two tasks with a multi-scale module to achieve competititve results.

4.3. Qualitative Results

Fig. 6 shows that our Stage-II framework generates realistic person images based on the text description and the predicted poses from Stage-I. In fact, the text guided image synthesis can be regarded as a semi-supervised problem, as there are only image-text pairs in the training dataset without ground-truth images corresponding to different text descriptions w.r.t the same identity. Nevertheless, by editing the descriptive words of various pedestrian appearance parts (*e.g.*, shirt, pants), our model is able to accurately change these parts in the image generation procedure. It indicates that our model is able to capture sufficient information from the text, while holding an effective control on the text.

Model	VQA score
mSIS	0.275
mAttnGAN	0.139
mPG^2	0.110
SAU	0.205
Ours	**0.334**

Table 2: VQA perceptual score (higher is better).

Our two-stage based method can edit both the pose and the attribute of the identity in the reference image by the natural language description, which is an interactive editing process for users. Fig. 7 shows the results of the predicted

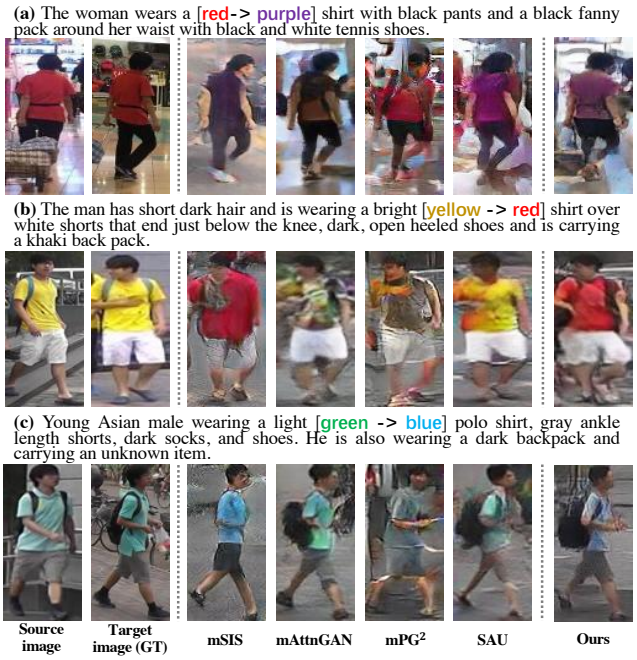

Figure 8: Qualitative comparison of our method and the baselines. Our method generates more valid and vivid person images.

poses and the generated images. Our method enhances the ability of the text for both the pose and the attribute interpolation. We can even change more than one word about the color attribute in the text description, and the synthesis is reasonable and correct corresponding to the text.

Fig. 8 shows a qualitative comparison of our method and different baselines. In the experiment, we find that there is a trade-off among identity invariance, pose transfer, and attribute transfer. For instance, mPG^2 first changes the pose, and then transfers the attribute of the person. However, the better the pose changes, the more difficult mPG^2 is to transfer the attribute. It is mainly because the distribution of the optimal pose is different from the distribution of the optimal attribute. This is also pointed out by [3] that the distinction of learned distribution may hurt the learning of feature representations when generating images. It is worth mentioning that although SAU adopts a single attentional module, its results are relatively better than those of the other baselines. However, SAU only integrates embedded features of different modalities at the smallest scale. In the experiment, we observe that this leads to more unreal attribute transfer. Thus, we use $m = 3$ attentional upsampling in our complete framework. Intuitively, the multi-scale upsampling module exploits receptive fields of different ranges to learn visual-word mapping on diverse spatial scales so as to better generate more realistic details. (*e.g.*, the woman's pack is preserved by our method in Fig. 8(a).)

Pose inference using GANs. Fig. 9 shows the selected basic poses and the inferred poses given text descriptions. The inferred poses are largely different from the basic poses, and our Stage-I model is able to concentrate on specific key words in the text (*e.g.*, walking, carrying) as these key words imply large changes in the posture of specific body parts (*e.g.*, arms, legs). Our model learns to adjust these details so that the inferred poses are much closer to the real ones, providing precise target poses for subsequent procedure of person image generation. We also implement a coordinate regression method as the baseline. As shown in Fig. 9, the coordinate regression method may lead to the distortion of some joints.

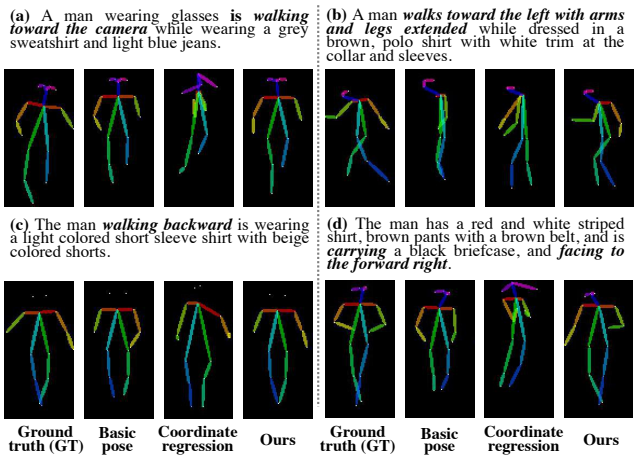

Figure 9: Qualitative comparison between our text guided pose generator and the coordinate regression method. The coordinate regression method may result in some distortions of joints, and our text guided pose generator generates more reasonable poses.

5. Conclusion

In this paper, we present a novel two-stage pipeline to manipulate the visual appearance (pose and attribute) of a person image based on natural language descriptions. The pipeline first learns to infer a reasonable target human pose based on the description, and then synthesizes an appearance transferred person image according to the text in conjunction with the target pose. Extensive experiments show that our method can interactively exert control over the process of person image generation by natural language descriptions.

Acknowledgments. This work is supported in part by NSFC (61672456), Zhejiang Lab (2018EC0ZX01-2), the fundamental research funds for central universities in China (No. 2017FZA5007), Artificial Intelligence Research Foundation of Baidu Inc., the Key Program of Zhejiang Province, China (No. 2015C01027), the funding from HIKVision, and ZJU Converging Media Computing Lab.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 1, 2

[2] 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 *CVPR*, 2018. 2

[3] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017. 8

[4] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *ICCV*, 2017. 2, 6

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 1, 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[7] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007. 6

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[9] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2

[10] 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

[11] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017. 6

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[13] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2

[14] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 6

[15] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 6

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 6

[17] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 1, 2, 4, 6

[18] Liqian Ma, Jia Xu, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 1, 2, 5, 6

[19] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *NIPS*, 2018. 3, 6

[20] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016. 2

[21] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. *arXiv preprint arXiv:1712.02225*, 2017. 3

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1, 2

[23] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016. 2

[24] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 5, 6

[25] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016. 2

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 5, 6

[27] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 6

[28] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 2018. 1, 2, 5, 6

[29] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017. 1, 2

[30] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017. 1, 2

[31] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. 2

[32] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018. 1

[33] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. Person re-identification meets image search. *arXiv preprint arXiv:1502.02171*, 2015. 6

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*, 2017. 2

[35] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017. 2