Learning Realistic Human Reposing using Cyclic Self-Supervision with 3D Shape, Pose, and Appearance Consistency

Soubhik Sanyal[2*]　　　Alex Vorobiov[1]　　　Timo Bolkart[2]　　　Matthew Loper[1]

Betty Mohler[1]　　　Larry Davis[1]　　　Javier Romero[1]　　　Michael J. Black[1]

[1]Amazon　　　[2]Max Planck Institute for Intelligent Systems, Tübingen

{ssanyal,tbolkart}@tue.mpg.de　　{vorobioo,mloper,btmohler,lrrydav,javier,mjblack}@amazon.com

Figure 1: **SPICE** (Self-supervised Person Image CrEation) generates an image of a person in a novel pose given a source image and a target pose. Each triplet in the figure consists of the source image (left), a reference image in target pose (middle) and the generated image in the target pose (right); input and reference images are from the DeepFashion test set [22].

Abstract

Synthesizing images of a person in novel poses from a single image is a highly ambiguous task. Most existing approaches require paired training images; i.e. images of the same person with the same clothing in different poses. However, obtaining sufficiently large datasets with paired data is challenging and costly. Previous methods that forego paired supervision lack realism. We propose a self-supervised framework named SPICE (Self-supervised Person Image CrEation) that closes the image quality gap with supervised methods. The key insight enabling self-supervision is to exploit 3D information about the human body in several ways. First, the 3D body shape must remain unchanged when reposing. Second, representing body pose in 3D enables reasoning about self occlusions. Third, 3D body parts that are visible before and after reposing, should have similar appearance features. Once trained, SPICE takes an image of a person and generates a new image of that person in a new target pose. SPICE achieves state-of-the-art performance on the DeepFashion dataset, improving the FID score from 29.9 to 7.8 compared with previous unsupervised methods, and with performance similar to the state-of-the-art supervised method (6.4). SPICE also generates temporally coherent videos given an input image and a sequence of poses, despite being trained on static images only.

1. Introduction

Given a single source image of a person, can we generate a realistic image of what they would look like from a different viewpoint, in a different pose? While this problem is inherently ambiguous, there is significant statistical regularity in human pose, clothing, and appearance, that could make this possible as illustrated in Fig. 1. A solution to the problem would have widespread applications in online fashion, gaming, personal avatar creation or animation, and has consequently generated significant research interest [6, 16, 31, 33, 38, 42].

---

*This work was done during an internship at Amazon.

Recent work focuses on generative modeling [8, 13, 15, 49], especially using conditional image synthesis. One set of methods uses supervised training [6, 20, 21, 32], which requires paired training images of the same person in different poses with the same appearance and clothing. Requiring such paired data limits the potential size of the training set, which can impair robustness and generalization. Consequently, we address this problem *without any paired data* by developing a self-supervised approach. Such self-supervised formulations have also received significant recent attention [7, 26, 30, 42].

Our novel formulation builds on the idea of cycle-consistency [49] with some important modifications. For the forward direction of the cycle, the method takes a source image, source pose and target pose and generates a target image conditioned on pose and appearance. The reverse direction takes this generated image and regenerates the source image by switching the source and target conditions. The goal is to minimize the difference between the original input image and the one synthesized through the cycle. The problem is that this approach can have a trivial solution in which the cycle produces the identity mapping. To address this, previous methods [30, 38] constrain the target image generation with 2D information. Human bodies, however, are non-rigid 3D entities and their deformations and occlusions are not easily expressed in 2D. We show how leveraging 3D information, automatically extracted from images, constrains the model in multiple ways.

Specifically, our method, called SPICE (Self-supervised Person Image CrEation), exploits the estimated 3D body to constrain the image generation, enabling self-supervised learning. In particular, we estimate the SMPL body model [23] parameters corresponding to both the input and the generated target image. Since the input and target image only differ in terms of their pose, their body shape should be the same. SMPL makes this easy to enforce because it factors body shape from pose. Using this we introduce two losses. First, we use a pose loss that encourages the body pose in the generated image to match the target pose in 3D. Second, we add a shape consistency loss that encourages the person in the generated image to have the same 3D shape as the person in the source image (Fig. 2).

These two constraints, however, are not sufficient to generate images with the correct appearance since they only force the model to generate an image with the right shape and pose. There is no constraint that the generated image has the appearance of the source image (e.g. clothing, hair, etc.). Prior work addresses this by enforcing a perceptual loss between patches at each 2D joint [30]. This is not sufficient when the body is seen with large viewpoint changes or where a body part becomes occluded; see Fig. 3. We solve this problem by introducing pose-dependent appearance consistency on the body surface instead of at the joints.
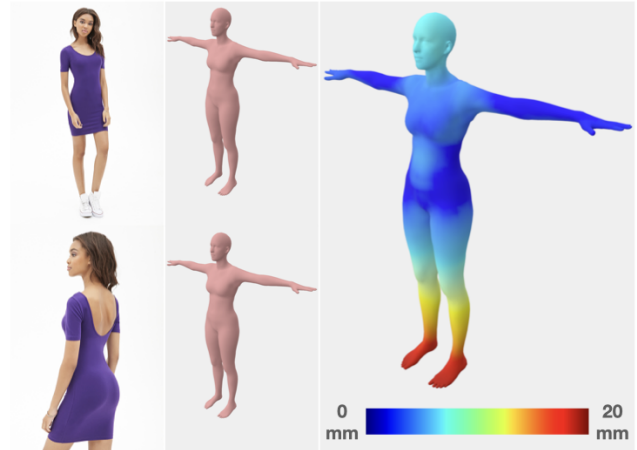

Figure 2: **Shape consistency:** The first column shows two images of the same person in two different poses and views. The second column shows the 3D bodies predicted by our 3D regressor and posed in a T-pose. The estimated 3D body shape is similar for the same subject across poses and views. The third column shows the per-vertex difference of both meshes, color coded from blue (0 mm) to red (20 mm).

The idea is that the projected surface of the 3D body in two different poses must have similar appearance features for matching parts of the body and this similarity should be weighted proportional to the relative global orientation difference between the 3D bodies.

In summary, we improve the realism of self-supervised human reposing by exploiting 3D body information in three novel ways: using a 3D pose loss, body shape consistency, and occlusion-aware appearance feature consistency. We train SPICE with our new constraints on unpaired data. Extensive experiments on the DeepFashion [22] and Fashion Video datasets [44] show the effectiveness of our model qualitatively and quantitatively. SPICE significantly outperforms the prior state-of-the-art (SOTA) un/self-supervised methods and is nearly as accurate as the best supervised methods.

2. Related Work

Methods for reposing images of humans can be broadly divided into two categories: supervised or unsupervised. While both approaches rely on generative modeling [8, 13, 15, 49], the supervised methods require paired ground truth: source and target training images of the person in different poses. Our approach falls into the unsupervised or self-supervised category, in which we do not use paired training data. We address each class of methods below.

Supervised methods: Supervised approaches learn to transform a source image given a source pose and a target pose [2, 5, 6, 9, 16, 19–21, 24, 25, 27, 28, 31, 32, 34–36, 39–41, 43, 45, 50]. Supervision is provided by the target
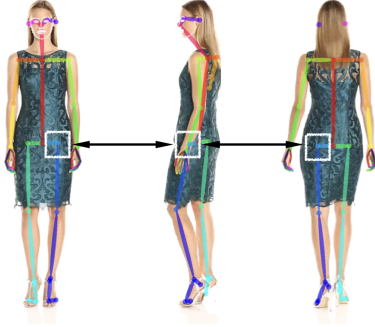
Figure 3: **Problem: patch loss based on 2D keypoints.** A person is seen in three different poses with the same clothing. A patch (white rectangle) is extracted at her left hip keypoint. Assuming that the appearance of the patch is the same across viewpoints is incorrect. Instead, SPICE uses the 3D body surface to reason about the regions of the body that are visible in multiple views. Keypoints are predicted by OpenPose [4] for this figure.

image during training and usually adversarial and perceptual losses are used to train the model [33]. The differences between methods usually lie in network inputs and their architectures. Dong et al. [5] synthesize the target image in two stages. First they generate a target pose segmentation from the source pose and use it in their soft-gated warping block architecture to render the person in the target pose. Knoche et al. [16] learn an implicit volumetric representation of the person to warp the source pose into the target pose. The volumetric representation is implicitly learned using an encoder decoder architecture. Li et al. [20] utilize a learned flow field to warp a person in a source pose to the target pose. The flow field is learned from 3D bodies and is used for warping at the feature level and pixel level in a deep architecture. Ma et al. [24] first generate a coarse image of the global structure of a human in the target pose from the source pose in a two stage network. This is then refined in an adversarial way in the second stage to get finer details. Sarkar et al. [32] compute a partial UV texture map using DensePose [10] from the source image. They use this as input to their network, which learns to complete the UV texture map and render it in a target pose using neural rendering. Siarohin et al. [35] propose a network architecture using deformable skip connections to tackle the problem. Tang et al. [39] propose a co-attention fusion model that fuses appearance and shape features from images, which they disentangle inside their architecture. They use two different discriminators for appearance and shape to jointly judge the generation. Zhu et al. [50] propose a progressive generator using a sequence of attention transfer blocks. Each of these blocks transfers certain regions it attends to and generates the image of the person progressively. Ren et al. [31] propose a new deep architecture where they combine flow-based operations with an attention mechanism.

Note that the above methods are all supervised and cannot be directly used in the self-supervised scenario. In contrast, our work is focused on unpaired data and we build on the Ren et al. [31] architecture to enable this. Thus our contribution does not lie with network architecture but, rather, in introducing novel constraints that make it possible to solve the problem without paired data.

Unsupervised or self-supervised settings: There is an increasing interest in solving the problem in an unsupervised/unpaired manner. Such approaches can work when paired data is not available or can increase robustness and generalization by combining paired and unpaired data. An early approach [26] divides the process in two stages. The first stage uses an auto-encoder-based architecture to learn the corresponding embedding space for pose, foreground and background from source images. The second stage maps Gaussian noise to the embedding space of pose, foreground and background and uses the pretrained decoder from the first stage to generate a person's image in a new pose. Yang et al. [42] train an appearance encoder from the source image to learn the appearance representation or embedding. They fuse the appearance embedding with the pose embedding coming from an image of a different person in different pose. In this way they generate the person's image in the new pose. Esser et al. [7] use a U-Net architecture conditioned on the output of a variational auto-encoder for appearance. They also try to disentangle pose and appearance of a person from the source images.

In general, the mentioned approaches attempt to disentangle shape, pose and appearance in the latent space from a 2D image, which is a hard problem. This results in a relatively poor image generation quality. Instead of learning this disentanglement from images we approach the problem differently. We extract the person's pose and shape information in a parametric decoupled 3D body representation, SMPL [23], and constrain our self-supervised generation. Furthermore, we also constrain appearance generation by leveraging the surface and projection of the 3D body.

Similar to our cyclic formulation, Pumarola et al. [30] and Song et al. [38] train their networks in a self-supervised CycleGAN [49] fashion. Additionally, [38] use semantic parsing maps as input to the network. They constrain their self-supervised generation with 2D information. We differ from these methods by constraining the self-supervised approach with 3D body information.

3. Method

SPICE requires a training dataset of tuples (I, P, R), each containing an image I of a person, their pose P as 2D keypoints, and a 2D rendering R. To generate R we fit the SMPL 3D mesh [23] to P using SMPLify [3], and render the mesh using a color wheel texture in UV space.

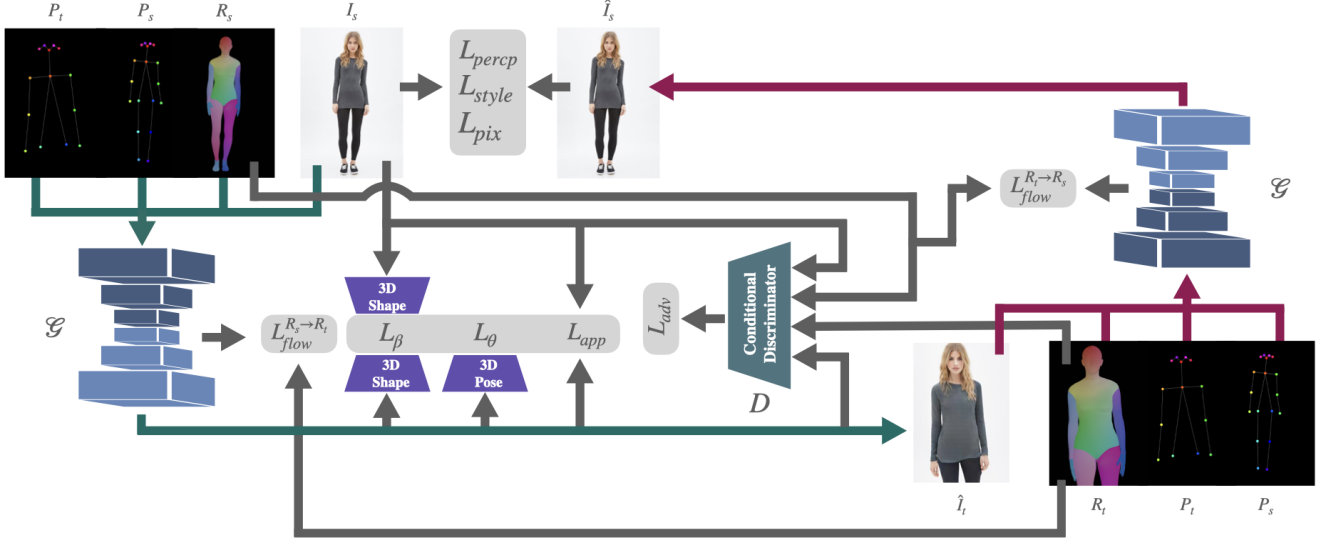We treat all the samples in the dataset as independent;

Figure 4: **Overview of SPICE:** Given a source image of a person I_s, source pose P_s, target pose P_t and 3D mesh rendering of the source pose R_s, the generator $\mathcal{G}$ generates a target image with the person in the target pose. Then the source and target pose are swapped and passed through $\mathcal{G}$ but with the generated target image as the source. This should re-generate the source image enabling the use of a cyclic self-supervision loss, $\mathcal{L}_{cycle}$, during training. To prevent trivial solutions, the cycle is constrained by losses on 3D pose $\mathcal{L}_{\theta}$, shape $\mathcal{L}_{\beta}$ and appearance $\mathcal{L}_{app}$, which are the main contributions of SPICE (Section 3), and an adversarial loss $\mathcal{L}_{adv}$. Note that the P_s and P_t are provided as input heat-maps to $\mathcal{G}$.

that is, our method does not require images of the same person wearing the same clothing in different poses (i.e. without direct supervision through paired data). During training, the source image I_s, source pose P_s, source rendering R_s, target pose P_t and target rendering R_t are given. SPICE then synthesizes the image $\hat{I}_t$, which is the reposed source image I_s, using a generator network $\mathcal{G}$ (Section 3.1):

$$\hat{I}_t = \mathcal{G}(I_s, P_s, R_s, P_t). \tag{1}$$

During training, we exploit cycle-consistency (Section 3.2). Specifically, we generate a synthetic version of the source image from $\hat{I}_t$ by reusing $\mathcal{G}$; i.e. $\hat{I}_s = \mathcal{G}(\hat{I}_t, P_t, R_t, P_s)$. This enables us to directly apply perceptual and pixel-wise losses between I_s and $\hat{I}_s$ to train $\mathcal{G}$. To prevent trivial solutions, we add 3D guidance (Section 3.3) and appearance constraints (Section 3.4) for $\hat{I}_t$. See Figure 4 for an overview of the SPICE training pipeline.

3.1. Generator architecture

Our generator $\mathcal{G}$ has two modules: a global flow field estimator and a local neural rendering module. The flow estimator module takes R_s, P_s, P_t as input and generates 2D warping fields at the feature level between the source and the target pose. The neural rendering module takes I_s and P_t as inputs and uses the generated warping fields at the feature level of its local attention blocks to generate $\hat{I}_t$. The loss for the flow estimator module can be written as,

$$\mathcal{L}_{flow} = \mathcal{L}_{flow}^{R_s \to R_t} + \mathcal{L}_{flow}^{R_t \to R_s}, \tag{2}$$

where $\mathcal{L}_{flow}^{x \to y}$ is the weighted addition of the sampling correctness loss and regularization loss for the generated flow fields, as proposed by Ren et al. [31]. Here, $\mathcal{L}_{flow}^{R_s \to R_t}$ is applied while synthesizing $\hat{I}_t$, and $\mathcal{L}_{flow}^{R_s \to R_t}$ is applied when regenerating $\hat{I}_s$ at the end of the cycle. The sampling correctness loss is the computed cosine similarity distance between the warped source features and target features. The source and target features come from a specific layer of a pre-trained VGG network [37] given the source and target renderings as input, respectively. The regularisation loss provides regularisation to the generated warping fields.

Our generator follows the design of Ren et al. [31] with the difference that the flow estimator is trained on source and target renderings (i.e. R_s and R_t), instead of source and target images (i.e. I_s and I_t), due to the unavailability of I_t in our setting. We refer to Ren et al. [31] for more details on the sampling correctness loss and the regularization loss.

3.2. Closing the cycle

Enforcing cycle consistency enables us to train SPICE with supervised losses between the source image I_s and the regenerated source image $\hat{I}_s$. Specifically, we minimize

$$\mathcal{L}_{cycle} = \lambda_{percep}\mathcal{L}_{percep} + \lambda_{style}\mathcal{L}_{style} + \lambda_{pix}\mathcal{L}_{pix}, \tag{3}$$

where the λ's are individual loss weights, and the perceptual loss $\mathcal{L}_{percep}$, style loss $\mathcal{L}_{style}$, [14], and pixel-wise loss

$\mathcal{L}_{pix}$ are defined as

$$\mathcal{L}_{percep} = \sum_j \left\| \phi_j(I_s) - \phi_j(\hat{I}_s) \right\|_1$$

$$\mathcal{L}_{style} = \sum_j \left\| \mathbb{G}(\phi_j(I_s)) - \mathbb{G}(\phi_j(\hat{I}_s)) \right\|_1$$

$$\mathcal{L}_{pix} = \left\| I_s - \hat{I}_s \right\|_1,$$

where ϕ_j is the activation map of the jth layer of a pre-trained VGG network [37], and $\mathbb{G}$ is the Gram matrix built from the activation map ϕ_j.

To generate realistic looking images, SPICE minimizes an adversarial loss by adding a discriminator, D, that discriminates between the fake images $\hat{I}_t$ and real images I_s. To provide pose information along with each image, we condition D on the corresponding rendering (i.e. R_t for $\hat{I}_t$, and R_s for I_s), by providing the concatenation of the two images as discriminator input. Formally, we minimize

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(\hat{I}_t, R_t))] + \mathbb{E}[\log D(I_s, R_s)]. \quad (4)$$

3.3. Pose and shape consistency

SPICE uses the SMPL [23] 3D human body model to enforce pose and shape consistency during training. SMPL combines identity-dependent shape blendshapes with pose-dependent corrective blendshapes and linear blendskinning (LBS) for pose articulation. Importantly, this formulation disentangles body shape form pose. Given parameters for shape $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$ and pose $\boldsymbol{\theta} \in \mathbb{R}^{3K+3}$, SMPL is a function, $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ that outputs a 3D mesh with $N = 6890$ vertices.

To extract SMPL shape and pose parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from I, we use a differentiable regressor [17], denoted as

$$\boldsymbol{\beta}, \boldsymbol{\theta} = f_{3D}(I). \quad (5)$$

Given the extracted SMPL parameters $\hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\theta}}_t = f_{3D}(\hat{I}_t)$, we define a loss that encourages the 3D rotation of the joints in the synthetic image, $\hat{\boldsymbol{\theta}}_t$, to be the same as the rotation of the joints in the target pose $\boldsymbol{\theta}_t$:

$$\mathcal{L}_{\boldsymbol{\theta}} = \left\| \boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t \right\|_1, \quad (6)$$

where $\boldsymbol{\theta}_t$ is obtained by running SMPLify [3] on P_t.

SPICE also enforces body shape consistency (Figure 2) based on the observation that while I_s and $\hat{I}_t$ differ in pose, their body shapes $\boldsymbol{\beta}_s$ (i.e. $\boldsymbol{\beta}_s, \boldsymbol{\theta}_s = f_{3D}(I_s)$) and $\hat{\boldsymbol{\beta}}_t$ must be the same, enforced by

$$\mathcal{L}_{\boldsymbol{\beta}} = \left\| \boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_t \right\|_1. \quad (7)$$

3.4. Appearance feature consistency

The above losses constrain pose and shape in $\hat{I}_t$, but do not guarantee that the appearance of $\hat{I}_t$ remains consistent
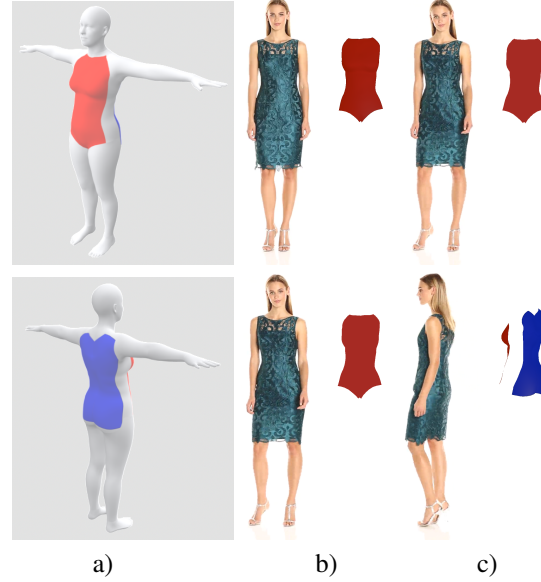

Figure 5: **Appearance feature consistency:** a) SMPL template with front (red) and back (blue) torso masks, b) and c) shows images of a person in different poses (left), and corresponding torso masks obtained by rendering the 3D body with the subject's pose. The appearance consistency loss is then applied on image segments for torso masks of the same color weighted by the relative pelvis rotation.

with I_s. Consequently, we formulate an additional constraint on the appearance of matching regions in I_s and $\hat{I}_t$ to be similar. Due to the unconstrained change in the pose between I_s and $\hat{I}_t$, we cannot apply the deep appearance loss (perceptual or style loss) directly between those images.

Instead, we leverage the 3D body mesh to apply an appearance loss between corresponding image segments. Given SMPL parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we render the mesh $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ with the texture from Figure 5 a) to get the image segments for the rendered front and back torso areas, shown in Figure 5 b), and c). Let M_{mask} denote a binary mask, with value 1 for pixels within the front/back torso segment and 0 elsewhere. Further, let $\mathcal{P}_{patch}$ denote $I \odot M_{mask}$, where $\odot$ is the Hadamard product. Both, m and p are cropped from M_{mask} and $\mathcal{P}_{patch}$, by the bounding box of the image segment.

Given image patches p_s and $\hat{p}_t$ together with binary masks m_s and $\hat{m}_t$, both extracted from I_s and $\hat{I}_t$, the appearance consistency is given as $l_{app} =:$

$$\lambda_{a_1} \sum_k \left\| \phi_k(p_s) \odot \psi_k(m_s) - \phi_k(\hat{p}_t) \odot \psi_k(\hat{m}_t) \right\|_1$$

$$+ \lambda_{a_2} \sum_j \left\| \mathbb{G}_j^\phi(p_s) \odot \psi_j(m_s) - \mathbb{G}_j^\phi(\hat{p}_t) \odot \psi_j(\hat{m}_t) \right\|_1 \quad (8)$$

where the λ's are weights, ϕ_k is the activation map of the

k^{th} layer of pretrained VGG network [37], $\mathbb{G}_j^\phi$ is the Gram matrix built from the corresponding activation map ϕ, and ψ is the down-sampling function for the corresponding layer.

Note that the appearance loss as it is formulated requires sufficient overlap of corresponding image features within the mask crop. We compute the appearance loss as:

$$\mathcal{L}_{app} = \lambda(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) \times l_{app}, \tag{9}$$

where $\lambda(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t)$ is a weighting function that depends on the relative pelvis rotation (i.e. rotation around the SMPL root joint) between the source and target pose:

$$\lambda(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) = \begin{cases} 1.0 & \text{if } 0° \leq |\boldsymbol{\theta}_s^{pel} - \boldsymbol{\theta}_t^{pel}| < 20° \\ 0.1 & \text{if } 20° \leq |\boldsymbol{\theta}_s^{pel} - \boldsymbol{\theta}_t^{pel}| < 40° \\ 0.01 & \text{if } 40° \leq |\boldsymbol{\theta}_s^{pel} - \boldsymbol{\theta}_t^{pel}| < 60° \\ 0.0 & \text{otherwise.} \end{cases} \tag{10}$$

3.5. Final loss

The total loss of the proposed approach is:

$$\begin{aligned} \mathcal{L}_{SPICE} = &\alpha_{cycle}\mathcal{L}_{cycle} + \alpha_{flow}\mathcal{L}_{flow} + \alpha_{adv}\mathcal{L}_{adv} \\ &+ \alpha_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\theta}} + \alpha_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}} + \alpha_{app}\mathcal{L}_{app}, \end{aligned} \tag{11}$$

where α_i are the corresponding loss weights. The following section provides details on how these weights are set.

4. Experiments

Datasets: SPICE is evaluated on two publicly available datasets, namely the DeepFashion In-shop Clothes Retrieval Benchmark [22] and the Fashion Video datatset [44]. The DeepFashion data are used for qualitative and quantitative comparisons and Fashion Video data for motion transfer examples, following Sarkar et al. [32]. The DeepFashion dataset [22] consists of 52712 high-resolution model images in fashion poses. The data are split into training and testing sets as in previous work [31, 50]. For training, we use 25341 images from the training set, in which body keypoints from nose to knee are at least visible. Further, 100 randomly selected images from the training set are held out as a validation set for model selection. The qualitative and quantitative evaluations are performed on the same 8570 image pairs as used by Ren et al. [31]. The Fashion Video dataset [44] consists of fashion pose video sequences of women in various clothing, captured with a static video camera. The dataset is split into 500 training and 100 test videos as done by Sarkar et al. [32], with each video containing roughly 350 frames. Please note that SPICE uses no paired images for training.

Training details: We use residual blocks as basic building blocks for $\mathcal{G}$. For more details of the architecture we refer the reader to Ren et al. [31]. We train SPICE with an image resolution of 256×256 for both datasets. We

Figure 6: **Qualitative results** on the Fashion Video dataset [44]. Video frames are synthesized from the source frame using the poses in the driving video. See **Sup. Mat.** for video and more examples.

DeepFashion	Unpaired	FID($\downarrow$)	LPIPS($\downarrow$)
Def-GAN [35]	✗	18.5	0.233
Pose-Attn [50]	✗	20.7	0.253
Intr-Flow [20]	✗	16.3	0.213
CoCosNet* [46]	✗	14.4	-
ADGAN ** [27]	✗	22.7	0.183
Ren et al. ** [31]	✗	6.4	0.143
VUNet ** [7]	✓	34.7	0.212
DPIG ** [26]	✓	48.2	0.284
PGSPT ** [38]	✓	29.9	0.238
SPICE (Ours)	✓	**7.8**	**0.164**

Table 1: Quantitative comparison of our method with other state-of-the-art methods. The * denotes that the method reports results for a different train/test split. The ** denotes that the metrics were recalculated using publicly available code and following the protocol described in **Sup. Mat.**; note that recalculation of the metrics results in different numbers from those reported in [31].

use spectral normalisation for both the generator and discriminator. The learning rate is $8e{-}4$ for $\mathcal{G}$ and $1.6e{-}3$ for the discriminator following similar GAN training strategy of Heusel et al [12]. We use 8 NVIDIA V100 GPUs

to train SPICE, where each GPU has a batch size of 8. We set the weights for different losses as follows: $\alpha_{cycle} = 1.0, \alpha_{flow} = 1.0, \alpha_{adv} = 1.0, \alpha_{\theta} = 0.01, \alpha_{\beta} = 0.01, \alpha_{app} = 1.0, \lambda_{a_1} = 0.01, \lambda_{a_2} = 10.0, \lambda_{percep} = 0.5, \lambda_{style} = 500.0, \lambda_{pix} = 5.0$. First, we train the flow-field estimator. Differing from Ren et al. [31], we use R_s and R_t together with keypoints due to the unavailability of I_t during training. R_s and R_t are used as the replacement for I_s and I_t respectively in their flow estimator module. We also finetune the 3D regressor f_{3D} on our training splits of the DeepFashion dataset [22] following a similar approach of Kolotouros et al. [18]. During the finetuning of f_{3D}, we use a similar representation proposed by Zhou et al. [48] for representing 3D rotations. Finally, we train the whole SPICE model end-to-end keeping the 3D regressor weights fixed. During a training iteration we use ROIAlign [11] to extract the desired regions from I_s and $\hat{I}_t$. We trained our models for 5 days (∼400 epochs). Inference for a single image takes 74 ms using a single NVIDIA V100 GPU.

Evaluation metrics: Following Ren et al. [31], we use Learned Perceptual Image Patch Similarity (LPIPS) [47] and FID [12] scores to evaluate our experimental results. LPIPS quantifies the perceptual distance between the generated image and the ground-truth image. The FID score is defined as the Wasserstein-2 distance between the distributions of real and generated images. We utilize the LPIPS score to evaluate the reconstruction error of SPICE, and the FID score to quantify the realism of the generated images. Image compression (e.g. JPEG) applied on reference or generated images significantly affects the FID scores. See **Sup. Mat.** for more details or [29] for a similar analysis. We also evaluated other metrics like contextual similarity [27] and object keypoint similarity [1] and provide the results in **Sup. Mat.** Details about the protocol for computing FID and LPIPS can also be found in **Sup. Mat.**

Quantitative evaluation: Table 1 quantitatively compares our method and other state-of-the-art approaches on the DeepFashion dataset [22]. We compare with Def-GAN [35], Pose-Attn [50], Intr-Flow [20], CoCosNet [46], ADGAN [27], Ren et al. [31], DPIG [26], VUNet [7] and PGSPT [38]. Note that Def-GAN, Pose-Attn, Intr-Flow, CoCosNet, ADGAN and Ren et al. are supervised methods, requiring the ground-truth image of a person in the target pose and clothing during training. In contrast, our method is unsupervised and comparable with the bottom half of the table (i.e. [7, 26]). We have used the publicly available code provided by Ren et al. [31], ADGAN [27], VUNet [7], DPIG [26] and PGSPT [38] to regenerate the images on our test split and recompute the metrics. SPICE achieves state-of-the-art results among unpaired methods and competitive results when compared with supervised methods.

Qualitative evaluation: Figure 1 shows results on the DeepFashion test split. SPICE does a good job of preserv-

Figure 7: **Qualitative comparison:** More results can be found in the **Sup. Mat.**

ing the target appearance and pose despite the large pose change. Figure 7 provides a qualitative comparison with other un/self-supervised methods on the DeepFashion test split. SPICE generates more realistic and high quality images while preserving pose and appearance compared with DPIG [26], VUNet [7] and PGSPT [38]. See **Sup. Mat. Video** for more visual results on the DeepFashion test split.

Motion transfer: If you can generate one pose, you can generate a sequence of poses. Consequently, we show video generation on the test split of the Fashion Video dataset in Fig. 6. We randomly select one video from the test split to act as a driving video that provides the target pose. We take the first frame of the other videos from the test split as the source image and generate the whole sequence from these. Please note that we did not train SPICE to generate videos; i.e. there is no video supervision or temporal consistency. See **Sup. Mat. Video** for examples of generated videos.

Ablation study: Table 2 summarizes our ablation study, which removes one loss at a time from the model. The configuration "SPICE w unconditional D" means that we give the generated image to the discriminator without conditioning on pose by concatenating the renderings. Our full model better preserves details, pose, and has better overall image quality. Trained without the pose loss, the generator has less information about the self-occlusions of the body. Therefore it tends to generate poses that are not possible for a real person, e.g. growing legs inside another leg, etc. If we train SPICE excluding the shape loss, the generator has less information about the 3D body shape of the person in the source image, which can lead to inconsistent deformations of shape in the generated images; e.g. having bigger hips with a very thin waist, etc. Excluding the appearance loss during training leads to less detailed reconstructions and an overall reduced clothing consistency. Figure 8 and

Configuration	FID($\downarrow$)	LPIPS($\downarrow$)
SPICE w/o shape loss	8.7	0.166
SPICE w/o pose loss	8.4	0.165
SPICE w/o appearance loss	9.9	0.164
SPICE w unconditional D	10.0	0.167
SPICE	7.8	0.164

Table 2: Ablation study on DeepFashion test set [22].

Figure 8: **Loss specific artifacts:** Each row shows artifacts when training without a specific loss. Top: without shape loss. Middle: without pose loss. Bottom: without appearance loss. From left to right: source image, reference image in the target pose, generated without the corresponding loss, and SPICE, respectively.

the **Sup. Mat. Video** illustrate such loss-specific artifacts.
Discussions and limitations: While the DeepFashion dataset [22] provides paired data, these pairs do not always have the same outfit, as can be seen in the bottom row of Figure 7. We manually checked 500 random sampled pairs of the training data, and found that in 16% of the pairs one of the images contains additional accessories or new clothes. This can be a burden for fully supervised methods. Instead, we take the extreme approach of pure self-supervision to see how far this can be pushed. For extreme pose/view changes, the solution is highly ambiguous: there is no way to know the front of an outfit from the back or vice versa. Although SPICE generates a plausible solution, the result might not match the real invisible details. A practical use case would limit the range of pose variation between source and target. SPICE requires the target im-

Figure 9: **Limitations:** SPICE has difficulty super-resolving fine details when zooming, dealing with extreme closeups, and generating humans from clothing images without humans.

age of a person where much of the body is in view (Fig. 9). Our model has difficulties preserving fine patterns when the camera zooms in Fig. 9. Zoom requires super-resolution which is a research topic in itself.

5. Conclusion

We have presented SPICE, a novel approach to repose clothed humans from a single image. SPICE is trained in a self-supervised fashion without paired training data by exploiting cyclic consistency. Our key insight is to use 3D body information during training in different ways to constrain the image generation. First, SPICE leverages a parametric 3D body model and a 3D body regressor to constrain body shape and pose. Second, SPICE uses the 3D body mesh to coherently segment source and generated images to enforce an occlusion-aware appearance feature consistency. Third, SPICE conditions a discriminator on colored mesh renderings to increase the quality of the generated images. Once trained, SPICE takes a single image and a target pose specified by 2D keypoints, and generates an image of the same person in the target pose. SPICE generates images that are significantly better than previous unsupervised methods, and that are similar in quality to the state-of-the-art supervised method. Additionally, SPICE can readily generate videos, although it is not trained for this task.

Adding 3D constraints to the reposing problem enables a number of applications that go beyond the scope of this paper and belong to future work. Although we used our shape and appearance losses to keep those traits constant, they could as well be used to control the output model appearance (e.g. changing the pattern of a T-shirt) or shape (e.g. changing the body proportions of the model).

Disclosure. While MJB is also an employee of the Max Planck Institute for Intelligent Systems (MPI-IS), this work was performed solely at Amazon where he is a part time employee. At MPI-IS he has received research gift funds from Intel, Nvidia, Adobe, Facebook, and Amazon. He has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH.

References

[1] Object keypoints similarity. https://cocodataset.org/#keypoints-eval. 7

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8348, 2018. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 3, 5

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 3

[5] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 474–484, 2018. 2, 3

[6] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[7] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8857–8866, 2018. 2, 3, 6, 7

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 2

[9] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12135–12144, 2019. 2

[10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 7

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 6, 7

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711. Springer, 2016. 4

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2

[16] Markus Knoche, István Sárándi, and Bastian Leibe. Reposing humans by warping 3d features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1044–1045, 2020. 1, 2, 3

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 5

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 7

[19] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *European Conference on Computer Vision (ECCV)*, pages 380–394. Springer, 2018. 2

[20] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3693–3702, 2019. 2, 3, 6, 7

[21] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5904–5913, 2019. 2

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2, 6, 7, 8

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multiperson linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 3, 5

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 406–416, 2017. 2, 3

[25] Liqian Ma, Zhe Lin, Connelly Barnes, Alexei A Efros, and Jingwan Lu. Unselfie: Translating selfies to neutral-pose portraits in the wild, 2020. 2

[26] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 99–108, 2018. 2, 3, 6, 7

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

[28] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *European Conference on Computer Vision (ECCV)*, pages 123–138, 2018. 2

[29] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On buggy resizing libraries and surprising subtleties in fid calculation. *arXiv preprint arXiv:2104.11222*, 2021. 7

[30] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8620–8628, 2018. 2, 3

[31] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 4, 6, 7

[32] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6

[33] Haoyue Shi, Le Wang, Wei Tang, Nanning Zheng, and Gang Hua. Loss functions for person image generation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 1, 3

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Enver Sanganeto, and Nicu Sebe. Appearance and pose-conditioned human image generation using deformable gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2019. 2

[35] Aliaksandr Siarohin, Enver Sanganeto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3416, 2018. 3, 6, 7

[36] Aliaksandr Siarohin, Enver Sanganeto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 4, 5, 6

[38] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6, 7

[39] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[40] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. Structure-aware person image generation with pose decomposition and semantic correlation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[41] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 2

[42] Hongtao Yang, Tong Zhang, Wenbing Huang, Xuming He, and Fatih Porikli. Towards purely unsupervised disentanglement of appearance and shape for person images generation. In *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*, pages 33–41, 2020. 1, 2, 3

[43] Lingbo Yang, Pan Wang, Chang Liu, Zhanning Gao, Peiran Ren, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Xiansheng Hua, and Wen Gao. Towards fine-grained human pose transfer with detail replenishing network. *IEEE Transactions on Image Processing*, 30:2422–2435, 2021. 2

[44] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 2, 6

[45] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 12749–12756, 2020. 2

[46] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 6, 7

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[48] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 7

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[50] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2347–2356, 2019. 2, 3, 6, 7