Unpaired Pose Guided Human Image Generation

Xu Chen Jie Song Otmar Hilliges
AIT Lab, ETH Zurich
{xuchen,jsong,otmarh}@inf.ethz.ch

Abstract

This paper studies the task of full generative modelling of realistic images of humans, guided only by coarse sketch of the pose, while providing control over the specific instance or type of outfit worn by the user. This is a difficult problem because input and output domain are very different and direct image-to-image translation becomes infeasible. We propose an end-to-end trainable network under the generative adversarial framework, that provides detailed control over the final appearance while not requiring paired training data and hence allows us to forgo the challenging problem of fitting 3D poses to 2D images. The model allows to generate novel samples conditioned on either an image taken from the target domain or a class label indicating the style of clothing (e.g., t-shirt). We thoroughly evaluate the architecture and the contributions of the individual components experimentally. Finally, we show in a large scale perceptual study that our approach can generate realistic looking images and that participants struggle in detecting fake images versus real samples, especially if faces are blurred.

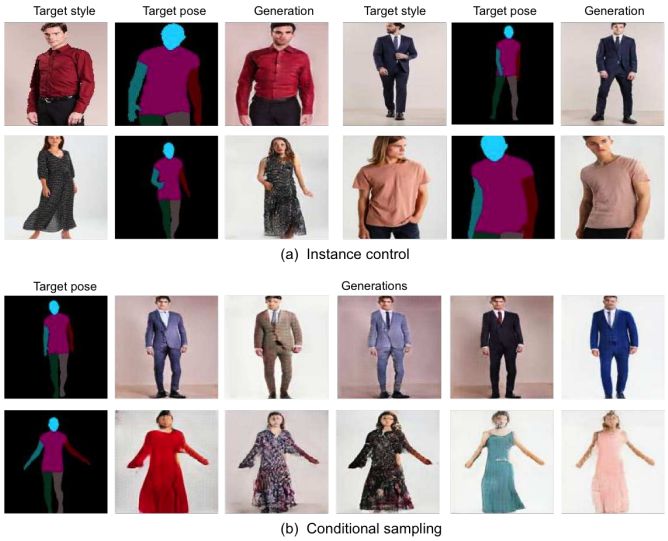

Figure 1: **Generating humans in clothing:** Our network takes a pose sketch as input and generates realistic images, giving users (a) instance control via an reference image, or (b) via conditional sampling, leading to images of variations of a class of outfits.

1. Introduction

In this paper we explore full generative modelling of people in clothing given only a sketch as input. This is a compelling problem motivated by the following questions. First, humans can imagine people in particular poses, wearing specific types of clothing – can machines learn to do the same? If so – how well can generative models perform this task? This clearly is a difficult problem since the input, a sketch of the pose, and the output, a detailed image of a dressed person, are drastically different in complexity, rendering direct image-to-image translation infeasible. The availability of such a generative model, would make many new application scenarios possible: cheap and easy-to-control generation of imagery for e-commerce applications such as fashion shopping, or to synthesize training data for discriminative approaches in person detection, identification or pose estimation.

Existing approaches typically cast the problem of generating people in clothing as two-stage paired image-to-image translation task. Lassner *et al*. [21] require pairs of corresponding input images, 3D body poses and so-called parsing images. While this allows for the generation of images with control over the depicted pose, such approaches do not provide any control over the type or even specific instance of clothing. However, many application scenarios would require control over all three.

In this paper we propose a new task that, to the best of our knowledge, isn't covered by any related work: we seek to generate images of people in clothing with (T1) control over the pose, and (T2) exact instance control (the type of clothing), or (T3) conditional sampling via class label, and with sufficient variance (e.g., blue or black suit). For example, provided an image of a person a model should be able to generate a new image of the person wearing a specific

clothing item (e.g., red shirt) in a particular pose (Fig 1 a). Or provided a pose and class label (e.g., dress) it should synthesize images of different variants of the clothing in target pose (Fig 1 b).

To tackle this problem, we furthermore contribute a new dataset, and a network architecture that greatly reduces the amount of annotations required for training which in turn drastically reduces human effort and thus allows for faster, cheaper and easier acquisition of large training data sets. Specifically, we only require *unpaired* sets of 3D meshes with part annotations and images of people in clothing. These sets only need to contain roughly similar distributions of poses, allowing for reuse of existing datasets.

Furthermore, our work contributes on the technical level in that we propose a single end-to-end trainable network under the generative adversarial framework, that provides active control over (i) the pose of the depicted person, (ii) over the type of cloth worn via conditional synthesis, and even (iii) the specific instance of clothing item. Note that in our formulation applying clothing is not performed via image-to-image transfer and hence allows for generation of novel images via sampling from a variational latent space, conditioned on a style label (e.g., t-shirt). Finally, our approach allows for the construction of a single inference model that makes the two-stage approach of prior work unnecessary.

We evaluate our method qualitatively in an in-depth analysis and show via a large scale user study ($n = 478$) that our approach produces images that are perceived as more realistic, indicated by a higher "fool-rate", than prior work.

2. Related Work

We consider the problem of generating images of people in clothing, with fine-grained user control. This is a relatively new area in the computer vision literature. Here we review the most closely related work and briefly summarize work that leverages deep generative models in adjacent areas of computer vision.

Generating people in clothing Synthesizing realistic images of people in different types of clothing is a challenging task and of great importance in many areas such as e-commerce, gaming and as potential source of training data for computer vision tasks. The computer graphics literature has dedicated a lot of attention to this problem including skinning and articulation of 3D meshes, simulation of physically accurate clothing deformation and the associated rendering problems [12, 7, 10, 29, 27, 20]. Despite much progress, generating photo-realistic renderings remains difficult and is computationally expensive.

Circumventing the graphics pipeline entirely, image-based generative modeling of people in clothing has been proposed as emerging task in the computer vision and machine learning literature.

One line of work [13, 38, 32, 43] targets so-called virtual try-on, transferring garment appearance from a high-quality photograph, to the corresponding body part(s) in the destination 2D image.

Another line of work, which is more related to ours, aims to generate human images with control over the exact body pose. [21, 23, 24, 35, 5, 4, 9, 30] generate images at the pixel level with variants of encoder-decoder architectures. [28, 36, 2, 8] further incorporate spatial warping to form human images in novel poses. However, the above methods rely on detailed annotations of natural human images, e.g. body poses [23, 24, 35, 5, 4, 9, 30, 28, 8, 36, 2] or clothing segmentations [21, 8], which are non-trivial to obtain. In contrast, our work aims to synthesize human images without these annotations. The only required training data for our network are *unpaired* sets of images of people in clothing and 3D meshes with part annotations. Being trained unsupervised, our method still provides control over both pose and appearance of the generated images.

Deep generative models Deep generative models that can synthesize novel samples have lately seen a lot of attention in the machine learning and computer vision literature. Especially generative adversarial networks (GANs) [11, 34, 31, 26] and variational autoencoders (VAEs) [19] have been successfully applied to the task of image synthesis. A particular problem that needs to be addressed in real tasks is that of control over the synthesis process. In their most basic form neither GANs nor VAEs allow for direct control over the output or individual parameters. To address this challenge a number of recent studies have investigated the problem of generating images conditioned on a given image or vector embeddings [18, 16, 41] such that they predict $p(\vec{y}|\vec{x})$, where $\vec{x}$ is a reference image or similar mean of defining the desired output (e.g., one-hot encoded class labels).

Image-to-image translation Generating people in clothing given a target pose can be regarded as an instance of image-to-image translation problem. In [16], automatic image-to-image translation is first tackled based on conditional GANs. While conditional GANs have shown promise when conditioned on one image to generate another image, the generation quality has been attained at the expense of lack of diversity in the translated outputs. To alleviate this problem, many works propose to simultaneously generate multiple outputs given the same input and encourage them to be distinct [45, 6, 3].

In [45], the ambiguity of the many-to-one mapping is distilled into a low-dimensional latent vector, which can be randomly sampled at test time and encourages generation of more diverse results. To train all of the above methods, paired training data is needed. However, for many tasks, paired information will not be available or will be very expensive to obtain. In [44], an approach called CycleGAN to translate an image from a source domain X to a target

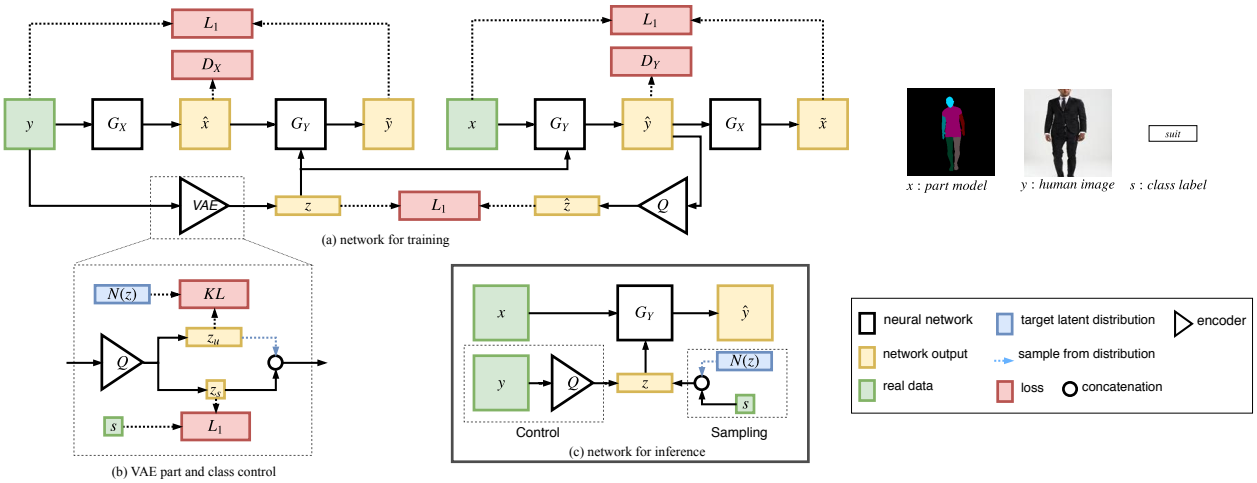

Figure 2: **Schematic architecture overview.** (a): network pipeline for training; (b): inset shows encoder in detail, consisting of a VAE and class label code; (c) the inference network generates synthetic images $\hat{y}$, given part model x as input and using either a style class label s or style depiction y to condition the generation process.

domain Y in the absence of paired examples has been introduced. Similar approaches have been proposed to enforce domain translation with cycle consistency loss [17, 42].

In our paper, we build on this prior work but leverage a formulation that gives control over individual dimensions when generating images and the capability to generate entirely novel samples via incorporation of a variational autoencoder that can be conditioned on a particular pose. A couple of concurrent works also recognize these limitations and propose architectures for the somewhat easier task of unsupervised multi-modal image domain translation [1, 15, 22]. We show that our architecture yields comparable results even though this task is not the main focus of our work.

3. Method

To conditionally synthesize images of people in clothing, while providing user control over the generation process, we propose a network combining the benefits of generative adversarial networks with cycle consistency and variational autoencoders in a single end-to-end trainable network. The aim is to generate high-quality images and to allow for the conditional generation of samples where the appearance is controlled via either an image (e.g., a specific suit) or a class label indicating the style of clothing (e.g., suits in general). Importantly our architecture can be trained via unpaired training data. That is 3D poses and real images do not need to correspond directly. The architecture, illustrated in Fig. 2 and dubbed Unpaired Pose-Guided GAN (UPG-GAN) combines elements of the GAN and VAE framework with several novel consistency losses to enable the fine-grained appearance control. In this section, we briefly introduce the basics of CycleGANs and VAEs and then detail our proposed architecture and training scheme.

3.1. Preliminaries: CycleGAN and VAE

To enable the desired unpaired training scheme, we build our framework on the basis of CycleGAN [44]. In the CycleGAN framework two generators $G_Y : X \rightarrow Y$ and $G_X : Y \rightarrow X$ translate images from one domain to another and attempt to fool the corresponding discriminators D_Y and D_X, classifying samples into real and synthetic data. This is expressed via the following minimax game:

$$\min_{G_X, G_Y} \max_{D_X, D_Y} (\mathcal{L}_{GAN}(G_X, D_X, Y, X) + \mathcal{L}_{GAN}(G_Y, D_Y, X, Y)), \tag{1}$$

where $\mathcal{L}_{GAN}$ is the standard GAN loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_{x \sim p(x)}[\log D(x)] + \mathbb{E}_{y \sim p(y)}[\log (1 - D(G(y)))]. \tag{2}$$

To prevent G_X and G_Y from neglecting the underlying information of the input source images, a cyclic reconstruction loss $\mathcal{L}_{cyc}$ is added:

$$\begin{aligned}\mathcal{L}_{cyc}(G_X, G_Y, X, Y) = \mathbb{E}_{x \sim p(x)}[\|G_X(G_Y(x)) - x\|_1] \\ + \mathbb{E}_{y \sim p(y)}[\|G_Y(G_X(y)) - y\|_1].\end{aligned} \tag{3}$$

The overall objective of CycleGAN is then given by:

$$\begin{aligned}\mathcal{L}_{CycleGAN}(G_X, G_Y, D_X, D_Y) = \mathcal{L}_{GAN}(G_X, D_X, Y, X) \\ + \mathcal{L}_{GAN}(G_Y, D_Y, X, Y) \\ + \mathcal{L}_{cyc}(G_X, G_Y, X, Y).\end{aligned} \tag{4}$$

This is optimized by alternating between maximizing the generator and minimizing the discriminator objectives.

Furthermore, we would like to endow the model with

the ability to stochastically generate samples of natural human images $\hat{y}$ (to attain natural variation in appearance), given a specified pose. For this task, we leverage VAEs so that a latent code z can be sampled from a prior distribution $\mathcal{N}(z)$. In the VAE setting, $\mathcal{N}(z)$ is commonly chosen to be an isotropic normal distribution $\mathcal{N}(0,1)$ with zero mean and unit variance. Applying the re-parametrization trick the corresponding KL-Divergence ($D_{KL}[\mathcal{N}(\mu(y),\sigma(y)^2)\|\mathcal{N}(0,1)]$) is minimized during training to regularize the latent distribution to be close to $\mathcal{N}(z)$, and μ and σ are encoded via the image y:

$$D_{KL} = \frac{1}{2}\sum_{k}(\exp(\sigma(y)^2) + \mu^2(y) - 1 - \sigma(y)^2). \quad (5)$$

3.2. Unpaired Pose Guided GANs (UPG-GAN)

We now discuss (1) how to endow the base CycleGAN architecture with control over the specific instance of clothing (provided via a reference image) and (2) how to allow for conditional generation of a diverse set of images via sampling where the style is controlled via a class label s.

Instance control. To allow for control over the clothing in the output image, we first introduce an additional encoder $Q : Y \rightarrow Z$ (dashed box in Fig. 2 (a)) to extract information from the input human image y into a latent code z. This code then serves as a guidance to reconstruct a human image $\tilde{y} = G_Y(\hat{x}, z)$ from the generated synthetic part model $\hat{x} = G_X(y)$. If z is simply fed to G_Y the generation process of $\hat{y}$ suffers from mode-collapse due to information hiding. During training G_X embeds a nearly imperceptible, high frequency signal into $\hat{x}$. At inference time x, the input to G_Y, is void of the hidden signal and generation of $\hat{y}$ converges to a single mode. This problem is especially severe when one domain, in our case the human images Y, represents significantly richer and more diverse information than the other, in our case the part models X (see Fig. 2). To prevent this, $\hat{y}$ is enforced to reflect the information encoded in z via introduction of a latent code consistency loss:

$$\mathcal{L}_c(G_Y, Q) = \mathbb{E}_{x\sim p(x), y\sim p(y)}[\|Q(G_Y(x,z)) - z\|_1]. \quad (6)$$

Here the encoder Q produces a style code $\hat{z} = Q(\hat{y})$ from $\hat{y}$.

Conditional sampling. In cases for which only the type of clothing matters but diverse image generation is desired (i.e., to triage different designs) we extend our architecture to accept a class label s for implicit guidance. To be able to generate varied images via sampling from a prior distribution, we propose a scheme similar to the VAE framework, albeit without an explicit decoder (see Fig. 2, inset). Conditioning with a specific type of clothing requires disentanglement of style-type information from other dimensions of the latent code such as color and texture of the clothing. Following the approach in ss-InfoGAN [37], we leverage easy to attain style class labels s while we do not provide labels for any of the other dimensions. The latent code z is then decomposed into a supervised part z_s, controlling the style class, and an unsupervised part $z_u \sim Q(z|y)$, where $z_s \cup z_u = z$. The following regularization term encourages the VAE's encoder E to learn a disentangled representation

$$\mathcal{L}_s(Q_s, Y, S) = \mathbb{E}_{y\sim p(y)}[\|Q_s(y) - s\|_1], \quad (7)$$

via enforcing similarity between z_s and its corresponding ground truth label s. The inference network (Fig. 2, c) can then be fed with labels s to control style classes while producing samples with sufficient variability (Fig. 1, b).

3.3. Training Schedule

Our overall objective function:

$$\mathcal{L}_{UPG-GAN} = \mathcal{L}_{GAN} + \mathcal{L}_{cyc} + \mathcal{D}_{KL} + \mathcal{L}_c + \mathcal{L}_s, \quad (8)$$

is optimized by the iterative training procedure given as pseudo-code in Alg.1. At the beginning of each iteration, one part model x and one human image y are randomly drawn from the corresponding datasets, and the ground truth label s for y is read, if available. The VAE encodes y to get the latent code z. Synthetic samples $\hat{x}$ and $\hat{y}$ are generated from x, y and z, and at the same time a latent code $\hat{z}$ is extracted from $\hat{y}$. The last step of the forward pass is to reconstruct $\tilde{y}$ from $\hat{x}$ and $\tilde{x}$ from $\hat{y}$ respectively. The overall loss is calculated and back-propagated through the network to update the weights.

Algorithm 1 Unpaired Pose Guided GAN

x part model, X part model dataset
y human image, Y human image dataset
s: class label (clothing), z: latent code
for *number of training epochs* **do**
  $x \leftarrow X, y, s \leftarrow Y$
  $z \leftarrow Q(y)$ where $z = z_s \cup z_u$
  $\hat{x} \leftarrow G_X(y)$, $\hat{y} \leftarrow G_Y(x, z)$
  $\tilde{x} \leftarrow G_X(\hat{y})$, $\tilde{y} \leftarrow G_Y(\hat{x}, z)$, $\hat{z} \leftarrow Q(\hat{y})$
  $\mathcal{L}^D_{GAN} \leftarrow (D(x)-1)^2+(D(y)-1)^2+D(\hat{x})^2+D(\hat{y})^2$

  $\mathcal{L}^G_{GAN} \leftarrow (D(\hat{x})-1)^2 + (D(\hat{y})-1)^2$
  $\mathcal{L}_{rec} \leftarrow \|\tilde{x} - x\|_1 + \|\tilde{y} - y\|_1$
  $\mathcal{L}_c \leftarrow \|\hat{z} - z\|_1$, $\mathcal{L}_s \leftarrow \|s - z_s\|_1$
  $\theta_{D_X}, \theta_{D_Y} \leftarrow Adam(\mathcal{L}^D_{GAN})$
  $\theta_Q \leftarrow Adam(\mathcal{L}^G_{GAN} + \mathcal{L}_{rec} + \mathcal{L}_s + \mathcal{D}_{KL})$
  $\theta_{G_X} \leftarrow Adam(\mathcal{L}^G_{GAN} + \mathcal{L}_{rec})$
  $\theta_{G_Y} \leftarrow Adam(\mathcal{L}^G_{GAN} + \mathcal{L}_{rec} + \mathcal{L}_c)$
end for

4. Experiments

Evaluating generative models is a difficult task since the main goal, that of synthesizing novel samples, implies that

Figure 3: Several images produced by ours and w/o $\mathcal{L}_s$, conditioned on target human images and poses. Each four images in a row form a group. From left to right in a group: target style image, target pose, our result and w/o $\mathcal{L}_s$.

no ground-truth information is available. Prior work on generating images of people in clothing has reported reconstruction accuracy. However, in our work this is not possible since we do not train on pairs of images. Furthermore, for the final task the two most important aspects are degree of control over the content and the final image quality. For these reasons we report mostly qualitative results but compare the various aspects of our proposed architecture in an ablative manner against the underlying building blocks (e.g., CycleGAN only). Finally, we report results from a large scale user study in which we asked participants to discriminate between randomly sampled real and fake images.

4.1. Dataset

One of the contributions in our work is the removal of the requirement for a task specific dataset. Training of the proposed architecture only relies on two separate sources of data: images of body part models and images of real people wearing varying types of clothing. To attain samples of *part models* we use the dataset of [21]. To attain the real *human images*, we crawled 1500 images from an online fashion store[1], including t-shirts, dress-shirts, dresses and suits. The label s was extracted from the online shop's categorization. To ensure rough correspondences between the body models and the images, we separated the datasets into those depicting full bodies and upper bodies only. Importantly these two data sources need not be directly paired and hence there is no need to fit the 3D body models to the 2D imagery. The dataset and code for network training and inference are released [2].

4.2. Implementation Details

We set the network input to a fixed size of 128×128 for computational reasons. The two generators G_X, G_Y (cf. Fig. 2) share the same architecture and we use a standard Downsampling-ResNetChain-Upsampling configuration. The VAE block and the classifier share the convolutional layers of a ResNet [14] architecture. During training, the learning rate is set to 0.00006. The weights of loss terms

[1] https://www.zalando.ch/

[2] https://github.com/cx921003/UPG-GAN

Figure 4: Conditional sampling comparison. (a): without $\mathcal{L}_s$ loss. (b): ours with $\mathcal{L}_s$.

are set to $\omega_s = 1$, $\omega_c = 10$, $\omega_{KL} = 0.01$, $\omega_{cyc} = 10$ and $\omega_{GAN} = 1$.

4.3. Ablation Study

To understand the effect of each component that we add to the base CycleGAN architecture, we conduct an ablation study. We contribute four novel aspects, namely the clothing encoder, the latent code consistency loss, the KL divergence loss and the class supervision loss. The clothing encoder and KL divergence loss are necessary to perform instance control (via source image) and conditional sampling (via class label) respectively. Therefore we focus our study on the latent code consistency loss and the class supervision loss. We train without the latent code consistency and its corresponding loss $\mathcal{L}_c$ and a network without $\mathcal{L}_s$ and evaluate their performance on both instance control and conditional sampling, compared to our full network. Note that we cannot compare directly to the base architecture (without both) since it has severe problem of mode collapse and also it fails to provide control over the depicted outfit.

Setup: During evaluation of instance control the realism of the generated human images and the similarity of style between the generation and the target are considered. To measure the realism we use a pre-trained faster-RCNN [33] to detect people in the generated images and report the detection accuracy and average confidence. To evaluate if the target style persists in the output, we use a pre-trained person re-identification network [40]. If the generated human image has a similar style to the target, the re-identification network should be able to re-identify it in the output image. To evaluate conditional sampling, we sample 20 human images for each input body part model and again compute the person detection accuracy and the average confidence. Moreover, we randomly pick 19 sample pairs for each input body part model, and measure the diversity of the generation with the average LPIPS score as proposed in [45].

Results: We first discuss the **instance control** results as shown in Tab. 1. As indicated by the re-identification con-

	Accuracy (control)	Confidence (control)	Re-ID (control)
w/o $\mathcal{L}_c$	61.4%	86.6%	0.382
w/o $\mathcal{L}_s$	90.7%	91.9%	0.642
Ours	**94.3%**	**94.4%**	**0.665**

Table 1: Ablation study for instance control.

	Accuracy (sample)	Confidence (sample)	Diversity (sample)
w/o $\mathcal{L}_c$	61.4%	86.4%	0.0001
w/o $\mathcal{L}_s$	80.7%	81.3%	**0.1240**
Ours	**93.9%**	**94.2%**	0.0700

Table 2: Ablation study for conditional sampling.

fidence, we can see that without the latent code consistency (w/o $\mathcal{L}_c$) the network cannot obey the target style image well. The classification loss $\mathcal{L}_s$ helps to slightly improve the style-preservation. In addition we can see that both the latent code consistency and the classification loss help to improve the realism, reflected by the decrease in detection accuracy and confidence when either of these two is absent. We can see from Fig. 3 that even though the version without $\mathcal{L}_s$ loss can produce satisfactory human images and also can keep the color correctly, it does not maintain the clothing type.

We now analyze the results in terms of **conditional sampling**. The low diversity score (0.0001) for the network trained without the latent code consistency loss $\mathcal{L}_c$ indicates that the provided latent codes are ignored during inference. By adding the latent code consistency loss, our full network can produce diverse samples with a much higher diversity score of 0.0700. Notably, the network without $\mathcal{L}_s$ achieves an even higher diversity score. However, this is due the unrealistic samples produced by this network, which is confirmed in the qualitative results and is also reflected in the low person detection scores. Fig. 4 shows that without $\mathcal{L}_s$ the network can produce diverse results but these are not consistent with the desired the type of clothing.

4.4. Latent Space Visualization

Fig. 5 visualizes the learned latent space, indicating that the latent codes indeed capture clothing information. We encode all of our training images that depict t-shirts into latent codes and project them to 2D via t-SNE [25] for visualization. The plot illustrates that latent codes cluster by clothing and texture, and not by pose, which indicates that pose and clothing are indeed disentangled.

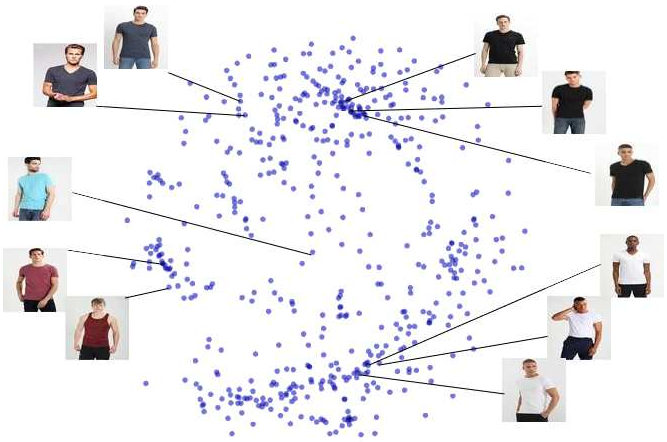

Figure 5: Visualization of the latent space.

Figure 6: Nearest training images to our generations.

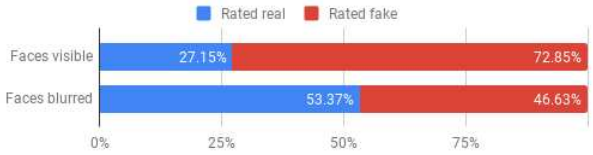

Figure 7: User ratings, without and with blurred faces.

4.5. Nearest Neighbor Analysis

A nearest neighbor analysis based on the structural similarity (SSIM) [39] metric, demonstrates that our learned model does not simply memorize the training data (see Fig. 6).

4.6. User Ratings

Finally, we assess the potentially most important metric to quantify the overall performance of the proposed approach. To better understand if the generated images look realistic we conduct a large-scale perceptual study. In this experiment we randomly sample images from the training dataset (*real*) and from synthetic samples generated by our model (*synth*). To isolate the influence of the facial region (which we do not treat specifically), we separate the participants in two groups where the first group judges images with full facial information and the second group judges im-

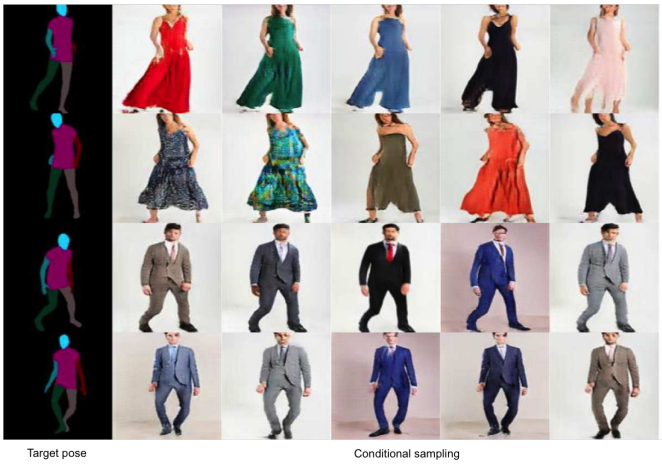

Figure 8: Samples drawn from the latent space z. Each row is conditioned one a particular pose and clothing style.

ages with blurred faces.

In total we asked $n = 478$ participants to judge 50 images (25 *real* / 25 *synth*). In a forced-choice setting, participants had to decide whether an image is real or synthetic. The participants did not know the true distribution of synthetic and real images and were not given any other instructions. Participants were recruited via mailing-lists. Fig 7 summarizes the results. With faces visible, synthetic images were rated as real with a fool-rate of $\sim 27\%$ which is significantly higher than previous work [21]. When removing the influence of the facial region via face blurring, this result improves to a fool-rate of $\sim 53\%$. This indicates that our model indeed synthesizes realistic samples and improves the perceived image quality over prior work.

4.7. Qualitative Results

As discussed in Section 3.2 and experimentally verified above, the proposed architecture can generate synthetic images, conditioned on either an example image of the target style or a class label. Fig. 8 illustrates that the architecture generates images of sufficient quality and is capable of producing samples with significant intra-class variation. Fig. 9 illustrates both concepts via a number of example sequences where each row shows images generated conditioned on a target pose and a specific clothing style. The samples contain significant diversity in both color and texture, while adhering to the pose guidance given by the input. In terms of image quality, we can see that the synthetic samples accurately capture the body pose information and generally appear realistic. The main challenge stems from the facial regions where neither the part model nor the class label provides any guidance and simultaneously the training data contains a lot of variation. One possibility to alleviate

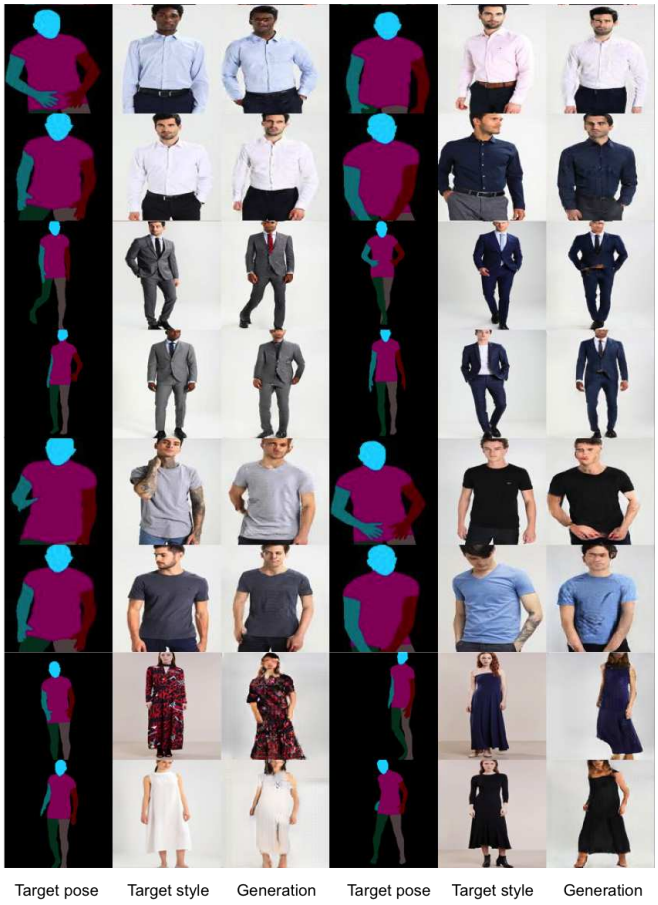

Figure 9: More examples on instance control: provided an image of a person to generate a new image of the person wearing this specific clothing in a particular pose

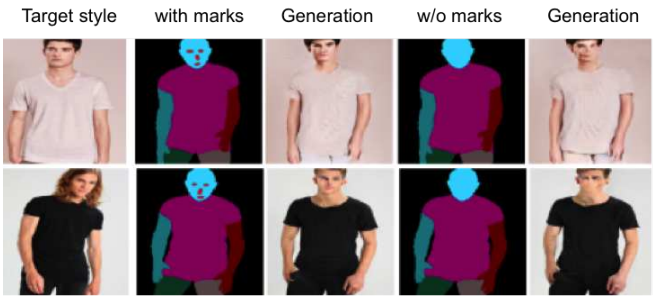

Figure 10: Improving synthetic faces via landmarks.

this issue could be to employ facial landmark annotations as shown in Fig. 10.

4.8. Limitation and Failure Cases

Fig. 12 depicts several challenging examples and failure cases. The main reason for unnatural synthetic images are uncommon or entirely unseen poses or viewing angles. Due

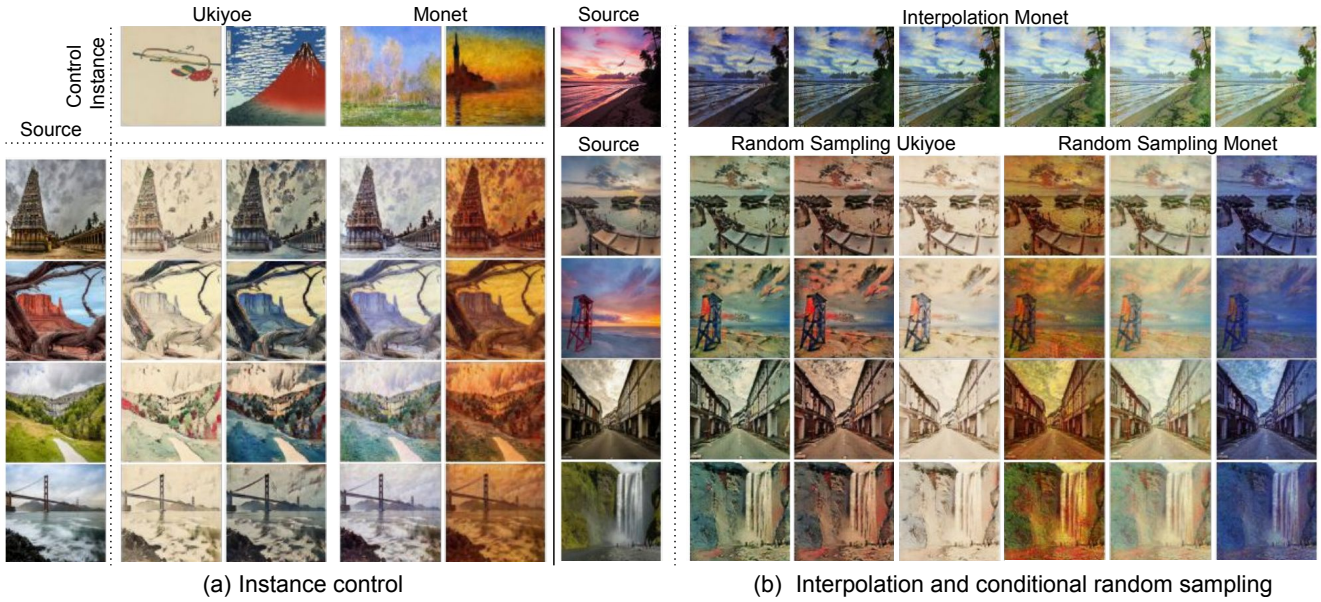

Figure 11: Qualitative results of photo2painting.

Figure 12: Failure cases due to uncommon poses or viewing angles.

to the unpaired nature of the proposed training regime, this kind of problem may be alleviated via the addition of more diverse training data.

4.9. Multi-model Conditional Image Translation

The focus of our work is generative modelling of human images. However, the architecture is general and can also be utilized in other application domains that need to synthesizing images, controlled by guidance that is specified in a different domain. Recent work has tackled the related problem of multi-modal image translation [15]. We demonstrate in Fig. 11 a photo-to-painting example. We use two styles of paintings and real photos from the dataset proposed in [44]. Where [15] train a network per style, we train only one single network for style translation. Furthermore, we show that we can control the style *and* texture of our generated paintings by maintaining content and texture from the input while translating the style from the reference (11, a). Furthermore, we demonstrate the capability to sample from the latent space, varying the texture of the generated image within the desired style (Fig. 11, b). Neither of these two functionalities can be achieved with direct-image-to-image approaches such as CycleGAN.

5. Conclusion

In this paper we have introduced the task of purely pose-guided generative modelling of humans in clothing. This task is challenging because it goes beyond the traditional setting of image-to-image translation. Given only a sketch of the human pose as input, we seek to generate realistic looking images that accurately depict the desired pose, and either a specific outfit (red dress), or to generate a number of images that depict the pose and variations of a class of outfit (different dresses). We have contributed a novel architecture that can generate high-quality images in an unpaired setting, while providing either direct instance or class-level control over the depicted style of clothing. We have experimentally shown that the proposed architecture is capable of creating realistic images and an ablative study showed the contributions of the different components of the architecture. A large-scale user-study shows that the generated images are seen as convincing, especially if the faces are blurred. Finally, we have demonstrated that the architecture can also be leveraged for other multi-modal image translation tasks.

Interesting directions for future work include, handling of uncommon poses and generation of images under novel viewpoints. Furthermore, it would be interesting to extend the task to the temporal domain which would also require modelling of the dynamics of the human body and its interactions with the non-rigid deformation of clothing.

References

[1] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018. 3

[2] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348. openaccess.thecvf.com, 2018. 2

[3] A. Bansal, Y. Sheikh, and D. Ramanan. PixelNN: Example-based image synthesis. In *International Conference on Learning Representations*, 2018. 2

[4] R. D. Bem, A. Ghosh, A. Boukhayma, T. Ajanthan, N. Siddharth, and P. Torr. A conditional deep generative model of people in natural images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1449–1458. ieeexplore.ieee.org, Jan. 2019. 2

[5] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. Aug. 2018. 2

[6] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017. 2

[7] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 479–488. IEEE, 2016. 2

[8] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin. Soft-Gated Warping-GAN for Pose-Guided person image synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 474–484. Curran Associates, Inc., 2018. 2

[9] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866. openaccess.thecvf.com, 2018. 2

[10] R. Goldenthal, D. Harmon, R. Fattal, M. Bercovier, and E. Grinspun. Efficient simulation of inextensible cloth. In *ACM Transactions on Graphics (TOG)*, volume 26, page 49. ACM, 2007. 2

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[12] P. Guan, O. Freifeld, and M. J. Black. A 2d human body model dressed in eigen clothing. In *European conference on computer vision*, pages 285–298. Springer, 2010. 2

[13] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[15] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018. 3, 8

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 2

[17] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 3

[18] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014. 2

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015. 2

[21] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model for people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 5, 7

[22] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1808.00948*, 2018. 3

[23] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 406–416. Curran Associates, Inc., 2017. 2

[24] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. 2

[25] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 6

[26] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017. 2

[27] R. Narain, A. Samii, and J. F. O'Brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):152, 2012. 2

[28] N. Neverova, R. Alp Guler, and I. Kokkinos. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 123–138. openaccess.thecvf.com, 2018. 2

[29] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3178–3185. IEEE, 2012. 2

[30] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628. openaccess.thecvf.com, 2018. 2

[31] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 2

[32] A. Raj, P. Sangkloy, H. Chang, J. Lu, D. Ceylan, and J. Hays. Swapnet: Garment transfer in single view images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682. openaccess.thecvf.com, 2018. 2

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 5

[34] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 2

[35] C. Si, W. Wang, L. Wang, and T. Tan. Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 118–126. openaccess.thecvf.com, 2018. 2

[36] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416. openaccess.thecvf.com, 2018. 2

[37] A. Spurr, E. Aksan, and O. Hilliges. Guiding InfoGAN with Semi-Supervision. *ArXiv e-prints*, July 2017. 4

[38] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604. openaccess.thecvf.com, 2018. 2

[39] 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.*, 13(4):600–612, Apr. 2004. 6

[40] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 5

[41] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 2

[42] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017. 3

[43] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5391–5399. openaccess.thecvf.com, 2018. 2

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 3, 8

[45] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017. 2, 5