Style and Pose Control for Image Synthesis of Humans from a Single Monocular View

KRIPASINDHU SARKAR, Max Planck Institute for Informatics
VLADISLAV GOLYANIK, Max Planck Institute for Informatics
LINGJIE LIU, Max Planck Institute for Informatics
CHRISTIAN THEOBALT, Max Planck Institute for Informatics

Fig. 1. We present StylePoseGAN, i.e., a new approach for synthesising photo-realistic novel views of a human from a single input image with explicit control over pose and per-body-part appearance. We generate images of higher fidelity compared to the state-of-the-art methods, especially with fine appearance details such as faces and texture patterns. Our method enables several applications such as pose transfer, garment transfer, and head swap. All synthesised results in this figure are shown as obtained from our model without further post-processing.

Authors’ addresses: KRIPASINDHU SARKAR, Max Planck Institute for Informatics, ksarkar@mpi-inf.mpg.de; VLADISLAV GOLYANIK, Max Planck Institute for Informatics, golyanik@mpi-inf.mpg.de; LINGJIE LIU, Max Planck Institute for Informatics, lliu@mpi-inf.mpg.de; CHRISTIAN THEOBALT, Max Planck Institute for Informatics, theobalt@mpi-inf.mpg.de.

1 INTRODUCTION

Creating photo-realistic images and videos of humans under full control of pose, shape and appearance is a core challenge in computer animation with many applications in movie production, content creation, visual effects and virtual reality, among others. Achieving this with the established computer graphics toolchains is an extremely complex and time-consuming process. First, a high-quality 3D human geometry and appearance model is required either manually designed by skilled artists or captured with dense camera arrays. To make the model animatable, sophisticated rigging techniques need to apply, and some manual post-processing (e.g., skinning weight painting) is often required. After that, computationally-expensive global illumination rendering techniques are needed to render the model photo-realistically. Some works [Casas et al. 2014; Volino et al. 2014; Xu et al. 2011] attempted to avoid such sophisticated toolchains with image-based rendering (IBR) techniques. However, these methods still have sub-optimal rendering quality and limited control over the renderings and are often scene-specific.

Recently, great progress in learning-based approaches for image synthesis from a single monocular input view has been made...
Sec. 4. Due to the separate conditioning of two aspects in different methods such as StyleGAN [Karras et al. 2019, 2020] and human poses and view-points in a pre-captured database and then applies retrieval-based texture synthesis. [Casas et al. 2014; Volino et al. 2014] compute a temporally coherent layered representation of appearance in texture space. However, the synthesis quality is limited by the quality of the geometry and appearance human model. To address the limitations of classical rendering methods, recent works integrated deep learning techniques into the classical rendering pipelines. Some methods [Kappel et al. 2020; Kim et al. 2018; Liu et al. 2020a, b, 2020b; Sitzmann et al. 2019a, b; Zhang et al. 2020] and dynamic scenes and enables playback and interpolation [Gafni et al. 2020; Li et al. 2020; Lombardi et al. 2019; Park et al. 2020a; Pumarola et al. 2020; Raj et al. 2020; Sida Peng 2020; Tretschk et al. 2020; Wang et al. 2020; Xian et al. 2020; Zhang et al. 2020] is not straightforward to extend these methods to synthesise human images of the full body with explicit control. Moreover, most of them are scene-specific. In contrast, our StylePoseGAN builds upon deep generative models and can synthesise photo-realistic human images with explicit control over body pose and human appearance.

2 RELATED WORK

We next review related works in human rendering, deep generative models and human pose transfer.

2.1 Classical and Neural Rendering of Humans

Photo-realistic rendering of a real human using classical rendering methods heavily relies on a high-quality geometry and appearance human model. To achieve high quality, real individuals need to be captured with sophisticated scanners and reconstructed by 3D reconstruction techniques. To synthesise the human image in a new pose, some sophisticated rigging techniques need to be applied to rig the human model, and a mapping from body poses to pose-dependent appearance or geometry models need to be learned. [Xu et al. 2011] propose a method which first retrieves the most similar poses and view-points in a pre-captured database and then applies retrieval-based texture synthesis. [Casas et al. 2014; Volino et al. 2014] compute a temporally coherent layered representation of appearance in texture space. However, the synthesis quality is limited by the quality of the geometry and appearance human model. To address the limitations of classical rendering methods, recent works integrated deep learning techniques into the classical rendering pipelines. Some methods [Kappel et al. 2020; Kim et al. 2018; Liu et al. 2020a, b, 2020b; Sitzmann et al. 2019a, b; Zhang et al. 2020] and dynamic scenes and enables playback and interpolation [Gafni et al. 2020; Li et al. 2020; Lombardi et al. 2019; Park et al. 2020a; Pumarola et al. 2020; Raj et al. 2020; Sida Peng 2020; Tretschk et al. 2020; Wang et al. 2020; Xian et al. 2020; Zhang et al. 2020] is not straightforward to extend these methods to synthesise human images of the full body with explicit control. Moreover, most of them are scene-specific. In contrast, our StylePoseGAN builds upon deep generative models and can synthesise photo-realistic human images with explicit control over body pose and human appearance.

2.2 Deep Generative Models

Generative Adversarial Networks (GAN) have made remarkable achievements in image generation in recent years. A GAN model...
has two components: generator and discriminator. The core idea of a GAN model is to use a generator to synthesise a candidate image from a noise vector $z$ sampled from the distribution of training images and let the discriminator evaluate whether the candidate is real or fake. The two components are trained together until the image generated by the generator is realistic enough to fool the discriminator. The first GAN model was introduced by Goodfellow et al. [2014], which was only able to synthesise low-resolution images with limited quality. To improve the quality, SAGAN [Zhang et al. 2019] introduces a self-attention mechanism into convolutional GANs, which allows the generator synthesising details using cues from all feature locations and the discriminator checking features in global image space. BigGAN [Brock et al. 2018] makes multiple changes on SAGAN for quality improvement. ProGAN [Karras et al. 2018] demonstrates photo-realistic images of human faces in a high resolution of $1024 \times 1024$ by training the generator and discriminator progressively from low resolution to high resolution. Since these methods use a single latent vector $z$ to resemble the latent distribution of training data, they cannot disentangle different attributes in the images so have limited control over image synthesis. StyleGANs [Karras et al. 2019, 2020] approach this problem by mapping $z$ to an intermediate latent space $w$, which is then fed into the generator to control different levels of attributes. Although they provide more control on image synthesis, it is still not able to completely disentangle different semantically meaningful attributes and control them in the synthesis. Recent works [Tewari et al. 2020a,b] extend StyleGAN to synthesise face images with a rig-like control over 3D interpretable feature parameters such as face pose, expressions and scene illumination. GAN-control [Shoshan et al. 2021] employ contrasting learning to train GANs with an explicitly disentangled latent space for faces, which can control identity, age, pose, expression, hair colour and illumination. Compared to faces, synthesising the full human appearance with control of 3D body pose and human appearance is a much more difficult problem due to more severe 3D pose and appearance changes.

Conditional GAN (cGAN) is a type of GAN, which provides conditional information for the generator and discriminator. cGAN is useful for applications such as class conditional image generation [Mirza and Osindero 2014; Miyato and Koyama 2018; Odena et al. 2017] and image to image translation [Isola et al. 2017; Wang et al. 2018b]. Most works [Isola et al. 2017; Mirza and Osindero 2014; Park et al. 2019; Wang et al. 2018a,b] require paired data for fully-supervised training. pix2pix [Isola et al. 2017] and pix2pixHD [Wang et al. 2018b] learn the mapping from input images to output images. GauGAN [Park et al. 2019] focuses on image generation from segmentation masks and designs an interactive tool for users to control over semantic and style in the image synthesis. To tackle the setting that paired data are unavailable, some works [Bansal et al. 2018; Choi et al. 2018; Liu et al. 2017a; Yi et al. 2017; Zhu et al. 2017] learns the mapping between two domains based on unpaired data. CycleGAN [Zhu et al. 2017] introduces a cycle consistency loss to enforce the translation property, i.e., that the inverse mapping of a mapping of an image should result in the original image. We propose a method that extends StyleGANs [Karras et al. 2019, 2020] and can synthesise photo-realistic images of a full human body with explicit control over 3D poses and the appearance of each body part. A concurrent work [Lewis et al. 2021] is similar to our method, which proposed a pose-conditioned StyleGAN2 latent space interpolation for virtual try-on. In contrast to their method, 1) we represent appearance in a pose independent normalised space which makes part based conditioning easier, 2) we use an explicit appearance encoder to encode the part based appearance latent vector $z$ with a single forward pass instead of optimizing for $z$, 3) we do not use explicit supervision through segmentation masks as needed in [Lewis et al. 2021] for latent code optimisation.

### 2.3 Human Pose Transfer

The human pose transfer problem is defined as transferring person appearance from one pose to another [Ma et al. 2017]. Most approaches formulate it as an image-to-image mapping problem, i.e., given a reference image of the target person, mapping the body pose in the format of renderings of a skeleton [Chan et al. 2019; Kratzwald et al. 2017; Pumarola et al. 2018; Siarohin et al. 2018; Zhu et al. 2019], dense mesh [Grigor’ev et al. 2019; Kappel et al. 2020; Liu et al. 2020b, 2019; Neverova et al. 2018; Sarkar et al. 2020; Wang et al. 2018a; Yoon et al. 2020] or joint position heatmaps [Aberman et al. 2019; Ma et al. 2017, 2018] to real images. Ma et al. [2017] design a two-stage framework, which first generates a coarse image of the person in the reference image with the target pose and refines the coarse image with a UNet trained in an adversarial way. To better preserve the appearance from the reference image to the generated image, some methods [Liu et al. 2020b; Sarkar et al. 2020] first map the human appearance in the screen space to UV space and feed the rendering of the person in the target pose with the UV texture map into an image-to-image translation network. Thanks to the explicit control over the pose and per-body-part appearance, StylePoseGAN can perform not only pose transfer but also be used for garment transfer and identity exchange.

### 3 METHOD

Given a single image $I$ of a person, our goal is to synthesise a new image of the same person in a different target body pose. The overall idea can be summarised as follows. We first extract pose $P$ and appearance $A$ from $I$. Second, we encode pose and appearance to the tensor encoding $E$ and $z$, respectively where $E$ is a 3D tensor with spatial dimensions (height and width), whereas $z$ is a vector. We then reconstruct $I$ using a high-fidelity style-based generator with $z$ as the noise vector and $E$ as the spatial input. The pose and appearance are further disentangled by training with the source $(s) –$ target $(t)$ image pairs $(I_s, I_t)$ of the same person with different poses, where we use the appearance of source $A_s$ and the pose of the target $P_t$ to reconstruct $I_t$ in a fully supervised manner. Our method is summarised in Fig. 4, and is described in detail in the following.

#### 3.1 Pose and Appearance Extraction

We use DensePose [Guler et al. 2018] to detect the human pose $P \in \mathbb{R}^{H \times W \times 3}$ from the image $I$, and represent the appearance $A \in \mathbb{R}^{H_a \times W_a \times 3}$ with the partial texture map of the underlying SMPL mesh. Here, $H \times W$ is the resolution of the generated image (with $H$ and $W$ denoting its height and width, respectively), and $H_a \times W_a$...
We use a StyleGAN2-based generator \cite{Karras_2020}. The ResNet-101-based DensePose network provided by the authors is pre-trained on COCO-DensePose dataset, and outputs 24 part-specific \( U, V \) coordinates of the SMPL model. For easier mapping, the 24 part-specific UV-maps are combined to a single UV-texture map \( A \) in the format provided in the SURREAL dataset \cite{Varol_2017} through a pre-computed lookup table.

The normalised texture map provides a pose-independent appearance encoding of the subject, where each part (out of 24) is assigned a specific region or a set of pixels in \( A \). Therefore, we can perform part-specific conditioning for the parts \( p_i, i \in \{1, \ldots, 24\} \) by changing values at its corresponding region \( A[Sp_i] \). Here, \( Sp_i \subset \mathbb{R}^{H_a \times W_a} \) are the pixel locations of the part \( p_i \) in the normalised texture map, and \( A[\cdot] \) is the indexing operation. See Fig. 5, 6 and Sec. 3.6.2 for more details.

### 3.2 Pose and Appearance Encoding

**PNet.** We encode \( P \) by a fully convolutional network \( PNet \) comprising four downsampling residual blocks that produce the encoded pose \( E \in \mathbb{R}^{H/16 \times W/16 \times 512} \). Note that based on the design of the generator that is used subsequently in our architecture, we do not need an explicit pose encoder. However, an encoded tensor with smaller spatial dimension is more suited for the StyleGAN2 based generator. See Secs. 3.3 and 4.3 for more details.

**ANet.** The appearance encoder \( ANet \) has the same architecture as \( PNet \) in its initial part. In contrast to \( PNet \), its spatial activation volume is further passed through convolutional layers and, finally, a fully connected layer to produce the appearance encoding \( z \in \mathbb{R}^{2048} \). Despite of the common architecture in the initial layers, \( ANet \) and \( PNet \) do not share any weights.

### 3.3 Image Generation with a Style-based Generator

We use a StyleGAN2-based generator \cite{Karras_2020} \( GNet \) that combines the pose encoding \( E \) and the appearance encoding \( z \) to reconstruct the back the input image \( I = GNet(E, z) \). The original StyleGAN takes a constant tensor \( S_{input} \) with spatial dimensions (with a predefined \( height \times width \times channel \)) as input on which convolutions are performed. A separate latent noise vector that controls the generated image is passed through a mapping network, and its output \( w \) is used to modulate the weights of the convolution layers, see Fig. 3. We observe that the tensor \( S_{input} \) can be used to provide spatial condition to the generator, instead of keeping it constant. Therefore, our \( GNet \) takes \( E \) as input to the convolutional layers. The convolutional weights are demodulated using the encoded appearance \( z \) to finally reconstruct the image \( I \). It comprises four residual blocks and four upsample residual blocks that transform an input tensor of dimensions \( H/16 \times W/16 \times 512 \) to an RGB image of dimensions \( H \times W \times 3 \). The architectural design of \( GNet \) follows StyleGAN \cite{Karras_2020} including bilinear upsampling, equalised learning rate, noise injection at every layer, adjusting variance of residual blocks and leaky ReLU. See Fig. 4 for an overview of StylePoseGAN.

### 3.4 Pose-Appearance Disentanglement by Paired Training

When the images of the same person in different poses are available for training, we can use them to further disentangle the pose and appearance. Given the source and target image pair \((I_s, I_t)\) of the same subject, we extract the corresponding appearances and poses \( A_s, A_t, P_s, P_t \). Next, we encode the poses and appearances using \( PNet \) and \( ANet \) to obtain the encodings \( s, t, z_s \) and \( z_t \). Finally, we generate the images \( I'_s \) and \( I'_{s\rightarrow t} \) as

\[
I'_s = GNet(E_s, z_s), \quad \quad I'_{s\rightarrow t} = GNet(E_s, z_t).
\]

Here \( I'_{s\rightarrow t} \) is the generated image of the source appearance in the target pose, which can be directly supervised during the training.

### 3.5 Training Details and Loss Functions

Given the input pairs \((I_s, I_t)\) and the generated images \( I'_s, I'_{s\rightarrow t} \), the entire architecture is trained end-to-end for the parameters of \( PNet, ANet \) and \( GNet \). We optimise the following loss:

\[
\mathcal{L}_{total} = \mathcal{L}(I'_s, I_s) + \mathcal{L}(I'_{s\rightarrow t}, I_t) + \lambda_{patch}\mathcal{L}_{patch}(I'_{s\rightarrow t}, I_t). \tag{2}
\]

The total loss \( \mathcal{L}_{total} \) consists of reconstruction loss \( \mathcal{L}(\cdot) \) and patch co-occurrence loss \( \mathcal{L}_{patch}(\cdot) \). The reconstruction loss

\[
\mathcal{L}(I_{gen}, I_{gt}) = \lambda_{L1}L_1 + \lambda_{VGG}L_{IP} + \lambda_{face}L_{face} + \lambda_{GAN}L_{GAN}
\]

comprises the following terms:

- **Reconstruction loss (L1).** We use L1 distance as a reconstruction loss to force \( I_{gen} \) and \( I_{gt} \) to be close to each other:

\[
L_1 = |I_{gen} - I_{gt}|. \tag{4}
\]

- **Perceptual Reconstruction Loss.** We use a perceptual loss \cite{Johnson_2016} based on the VGG network to enforce perceptual similarity between generated and the ground truth image. It is defined as the difference between the activations on different layers of the pre-trained VGG network \cite{Simonyan_2015}.
Given an image $I$ of a person, we extract the pose $P$ and appearance $A$ using DensePose [Gueler et al. 2018]. We then encode the pose and appearance to encodings $E$ and $z$ such that $E$ is a tensor and $z$ is a vector. We finally condition a high-fidelity style-based generator with the extracted pose and appearance to reconstruct back $I$. The pose and appearance are further disentangled by training with image pairs $(I_s, I_t)$ of the same person with a different pose where we use the appearance of source $A_s$ and the pose of the target $P_t$ to reconstruct $I_t$. The entire pipeline is trained end-to-end in a fully supervised manner with image reconstruction loss and adversarial loss.

**Face Identity Loss.** We use a pre-trained Face Identity Network to enforce similarity of the facial identity between $I_{gen}$ and $I_{gt}$:

$$L_{face} = |N_{face}(I_{gen}) - N_{face}(I_{gt})|,$$  

where $N_{face}$ is the pre-trained SphereFaceNet [Liu et al. 2017b].

**Adversarial Loss.** We use an adversarial loss $L_{GAN}$ with a discriminator $D$ of the identical architecture as in StyleGAN2 [Karras et al. 2020]. Please refer to StyleGAN2 for further details.

In addition, we use Patch Discriminator Loss $L_{patch}$ with a patch co-occurrence discriminator $D_{Patch}$. $D_{Patch}$ is trained such that the patches in $I_{gen}$ can not be distinguished from patches in $I_{gt}$. A similar idea is used by [Park et al. 2020b] in an unsupervised setting. Please refer to [Park et al. 2020b] for the architecture of $D_{Patch}$.

The final objective $L_{total}$ is minimised w. r. t. the parameters of $P_{Net}$, $A_{Net}$ and $G_{Net}$, while maximised w. r. t. $D$ and $D_{Patch}$.

### 3.6 Inference

Once trained, our single trained model can be used for pose transfer, garment and attribute transfer, and interpolation in the learned manifold of appearance. In summary, we can control the pose encoding $E$ and the appearance encoding $z$ independently to accomplish a wide range of tasks.

#### 3.6.1 Pose Transfer

For pose transfer, our method takes a source appearance $A_s$ (that is extracted from a human image), and a target DensePose $P_t$ as input. The re-rendered image of the source person...
in the target pose is then obtained as
\[ I_{\text{swap}} = GNet(PNet(P_b), ANet(A_b)). \] (7)

#### 3.6.2 Garments and Parts Transfer

Once trained for pose transfer, our model can be used for garment transfer without any modification. Given a source body image \( I_b \) (with appearance \( A_b \)) and a target garment image \( I_g \) (with appearance \( A_g \)), the aim is to reconstruct the person in \( I_b \) with the garments in \( I_g \) as \( I_{\text{swap}} \). To achieve this task, we construct a hybrid appearance image \( A_{b\rightarrow g} \) with the garment specific regions in \( A_g \) and body-specific regions in \( A_b \), see Fig. 6. The generated image with the swapped garments is then \( I_{b\rightarrow g} = GNet(PNet(P_b), ANet(A_{b\rightarrow g})) \). Here GNet, PNet and ANet are the trained models for the task of pose transfer.

### 4 EXPERIMENTAL RESULTS

#### 4.1 Experimental Setup

We use the In-shop Clothes Retrieval benchmark of DeepFashion dataset [Liu et al. 2016] for our main experiments. The dataset comprises around 52k high-resolution images of fashion models with 13k different clothing items in different poses. We use the training and testing splits provided by [Siarohin et al. 2019]. To filter non-human images, we discard all the images where we could not compute DensePose, resulting in 38k training images and 3k testing images. We train our system with the resulting training split and use the testing split for conditioning poses. We also show the qualitative results of our method on Fashion dataset [Zablotskaia et al. 2019] that has 500 training and 100 test videos, each containing roughly 350 frames.

We train our model for the task of pose transfer with paired data, as described in Sec. 3.5. In our experiments, texture resolution \( H_d \times W_d \) is chosen to be 256 × 256, while the output resolution \( H \times W \) is chosen to be 256 × 256 and 512 × 512 depending on compared methods. The loss weights (Sec. 3.5) are set empirically to \( \lambda_{AL} = 1, \lambda_{VGG} = 1, \lambda_{face} = 1, \lambda_{GAN} = 1, \lambda_{patch} = 1 \). For training, we use ADAM optimizer [Kingma and Ba 2015] with an initial learning rate of 0.002, \( \beta_1 = 0.0 \) and \( \beta_2 = 0.99 \). After the convergence of the training, we use the same trained model for all the tasks, i.e., pose transfer, garment transfer, style interpolation and motion transfer.

#### 4.2 Pose Transfer

We perform the experiment of pose transfer using the DeepFashion dataset. Given a human image and a target pose, we re-render the human in the target pose as described in Sec. 3.6.1. Our qualitative results are shown in Figs. 7 and 8.

#### 4.2.1 Comparison with State of the Art

We compare our results with seven state-of-the-art methods: Coordinate Based Inpainting (CBI) [Grigor’ev et al. 2019], Deformable GAN (DSC) [Siarohin et al. 2019], Variational U-Net (VU-Net) [Esser et al. 2018], Dense Pose Transfer (DPT) [Neverova et al. 2018], Neural Human Re-Rendering (NHRR) [Sarkar et al. 2020], and ADGAN [Men et al. 2020] at the resolution 256 × 256 and show the qualitative results in Fig. 7. We train and evaluate our model with the training-testing split provided by Deformable GAN [Siarohin et al. 2019]. This split is also used by all the aforementioned pose transfer methods with the exception of ADGAN [Men et al. 2020]. Training the official implementation of ADGAN with our training split did not converge. Therefore, we provide here the results from their trained model in spite of the significant overlap of the testing pairs in their training.

It can be seen that our results show higher realism and better preserve the identity and garment details compared to the other methods. We observe that StylePoseGAN also faithfully reconstructs high-frequency details, such as textures and patterns in the garments, which was not captured by any of the competing methods.

We next perform a quantitative evaluation with a subset of the entire testing pairs – 176 testing pairs that are used in the exiting works [Grigor’ev et al. 2019; Sarkar et al. 2020]. The following two metrics were used for comparison:

- **Structural Similarity Index (SSIM)** [Zhou Wang et al. 2004]. SSIM has been widely used in the existing literature for the problem of pose transfer. However, this metric often does not reflect human perception. It is observed that smooth and blurry images tend to have better SSIM than sharper images [Neverova et al. 2018; Zhang et al. 2018].

- **Learned Perceptual Image Patch Similarity (LPIPS)** [Zhang et al. 2018]. LPIPS captures human judgment better than existing hand-designed metrics, making it the most popular and important metric to evaluate generated images. The quantitative results are shown in Table 1. We significantly outperform the existing methods on both metrics. In terms of LPIPS, we observe an improvement of 19% (from 0.164 to 0.133) over the previous best result (NHRR).

#### 4.2.2 User Study

We conduct a user study to assess the visual quality of the pose transfer results by CBI, NHRR and our method. We follow the user study methodology introduced in [Sarkar et al. 2020], i.e., the covers a wide variety of source and target poses, and the distribution of males and females roughly reflects the same distribution in the training dataset. Moreover, multiple queries contain artefacts for our method, which make the decisions difficult. For each sample, we ask the following questions:

(Q1) Which view looks the most like the person in the source image?

(Q2) Which view looks the most realistic?

(Q3) Which view preserves fine appearance details (e.g., texture patterns, wrinkles and other elements) better?
We prepare 46 questions asked in a web-browser user interface in a randomised order. Each question contains a real source image of a person and there pose transfer results (by CBI, NHRR and our method) in randomised order.

27 anonymous respondents have submitted their answers, and the results are summarised in Table 2. Our method ranks first and is preferred in all three question types by a significant margin compared to CBI and NHRR. CBI is preferred twice out of 144 cases, i.e., once in Q1 (“person similarity”) and once in Q3 (“fine details”), see Fig. 9 for the corresponding image sets. Note that even if CBI was selected two times, the remaining two questions to the same set of images were decided in favour of our method (e.g., CBI preserved...
Fig. 8. High-resolution results (512 x 512) of our method for pose transfer. The conditioning pose is shown on the bottom right for each generated image.
We perform the following ablation study to see the usefulness of the different loss terms, we perform the experiment \textit{w/o VGG+Face} and \textit{w/o patch co-occurrence} by removing VGG and face losses, and patch co-occurrence. We see with both qualitative and quantitative results that VGG and face identity losses are crucial for the final results.

4.3.2 Losses. To see the usefulness of the different loss terms, we perform the experiment \textit{w/o VGG+Face} and \textit{w/o patch co-occurrence} by removing VGG and face losses, and patch co-occurrence. We see with both qualitative and quantitative results that VGG and face identity losses are crucial for the final results.

Table 2. The summary of the user study.

| Q1 ("person similarity") | CBI | StylePoseGAN (ours) |
|---------------------------|-----|----------------------|
| 2.2%                      |     | 97.8%                |
| Q2 ("overall realism")   | 0%  | 100%                 |
| Q3 ("fine details")      | 2.2%| 97.8%                |

4.3.3 Training strategy. We often do not have a training dataset containing images of the same person with multiple poses. To address the issue, we propose the ablation experiment \textit{No-Pairs} where we do not use source-target pairs for the training. Given an image \(I_s\), this design reconstructs back \(I_s\) by its own appearance \(A_s\) and pose \(P'_s = GNet(PNet(P_s), ANet(A_s))\). In addition, we use a random pose from the training set \(P_t\), and use it to reconstruct \(I_{s-t} = GNet(PNet(P_t), ANet(A_s))\). However, we do not apply any reconstruction loss on \(I_{s-t}\) because of unavailability of \(I_t\) during the training. Instead, we apply patch co-occurrence loss \(L_{patch}\) between \(I_{s-t}\) and \(I_t\). Therefore, the loss objective of Eq. (2) is modified to

\[
L_{total} = L(I_s, I_s) + \lambda_{patch} L_{patch}(I'_s, I_t).\]  

This experiment accounts for the unsupervised version of our method. We also perform an experiment with the exact setting but with no additional patch co-occurrence loss. A similar method is proposed in swapping autoencoder [Park et al. 2020b]. In contrast to Park et al., this experiment explicitly uses normalised pose and normalised appearance as input which enables better control of the conditioning pose and appearance. This experiment also provides us with the model for garment transfer and has all the advantages of our full method.

The usefulness of the co-occurrence loss can be seen in this unpaired scenario. As observed in Fig. 10, \textit{No-Pairs} with co-occurrence loss performs well. However, it lacks texture details and adds spurious patterns when the poses are too different. \textit{No-pairs} without co-occurrence loss do not preserve appearance. The paired training in our full method version forces the generator to account for the missing texture in highly occluded areas.

4.4 Motion Transfer

Our method can be applied to each frame of a driving video to perform motion transfer. To this end, we keep the image of the source person fixed and use the pose from the actor of the driving
Fig. 10. **Results of our ablation study.** Here, **No-PEncoder** does not use pose encoder, but a generator requiring more memory; **w/o VGG&Face** and **w/o co-occur** do not use VGG/Face and co-occurrence loss, respectively; **No-Pairs** baselines do not use paired training. We observe that our full method performs better than the baselines in almost all cases. The importance of co-occurrence loss during unpaired training is discussed in Sec. 4.3.

Video in our system to create image animation. Note that we do not exploit any temporal context, and we create the video frame by frame. We show our result on Fashion Dataset [Zablotskaia et al. 2019] in Fig. 12 and our supplementary video. We observe that our method does not show the shower-curtain effect that is typically present in translation-network-based generators (such as Pix2Pix). Even with the errors and inconsistencies in DensePose on the testing frames, our method keeps the fine wrinkle patterns consistent and provides pose-dependent appearance changes. Note that these wrinkle patterns are learned implicitly by our method through the training data.

### 4.5 Garment and Part Transfer

As explained in Sec. 3.6.2, our part-based encoding enables our model to perform garment transfer without any further training, i.e., no explicit dataset of garments are needed for our garment transfer functionality. The results are shown in Fig. 13. We observe that StylePoseGAN faithfully reconstructs the garments and body details, and seamlessly generates coherent human images with swapped garments.

The same idea can be used to transfer any other body parts. To this end, we show an example of head-swap in Fig. 14. Here we use the hybrid appearance $A$ that contains the partial texture of body from one image and head from another image, and the pose from
We can interpolate between two appearance vectors to generate coherently dressed humans with the properties of both target appearances.

Fig. 11. **Latent space interpolation.** We can interpolate between two appearance vectors to generate coherently dressed humans with the properties of both target appearances.

\[
\mathbf{z}_{\text{inter}} = \mathbf{z}_1 t + \mathbf{z}_2 (1 - t), \quad t \in [0, 1],
\]

where \(\mathbf{z}_1, \mathbf{z}_2\) are the appearance encodings of two human images, resulting in images of coherently dressed humans with the properties of both images. Examples are shown in Fig. 11.

the body image to create the head swap effect. Note that the identity is well preserved even though some conditioning head images are side views.

### 4.6 Style Interpolation

We find that our latent space is smooth. Interpolating appearance features \(\mathbf{z}_{\text{inter}}\) according to

\[
\mathbf{z}_{\text{inter}} = \mathbf{z}_1 t + \mathbf{z}_2 (1 - t), \quad t \in [0, 1],
\]

5 **DISCUSSION**

**Limitations/Failure Cases.** Because we encode the appearance by fully connected layers to a single vector, the spatial semantics in the appearance is destroyed. Therefore, our method has difficulty in reconstructing highly entangled spatial features such as images
or text in the clothes (see Figs. 15 and 16). Even in those cases, our method performs better than translation-network-based methods such as NHRR because of the powerful nature of our generator.

6 CONCLUSION

We presented StylePoseGAN, i.e., a new method for synthesising photo-realistic novel views of humans from a single monocular image allowing for explicit control over the pose and appearance without the need for sophisticated 3D modelling. StylePoseGAN significantly outperforms the current state of the art in the perceptual metrics and achieves a high level of realism of the synthesised images in our experiments. The quantitative results are confirmed by a comprehensive user study, in which our results are preferred over competing methods in 142 cases out of 144. We conclude that this is due to improved support of fine texture details of the human appearance such as facial features, textures and shading. The ablative study has confirmed that all design choices are necessary for the best photo-realistic results. We thus believe that our method, which often produces deceptively realistic renderings, is an important step towards unconstrained novel view rendering of scenes with humans, opening up multiple avenues for future research.

REFERENCES

Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Deep Video-Based Performance Cloning. Computer Graphics Forum 38, 2 (2019), 219–233.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2Shape: Detailed Full Human Body Geometry from a Single Image. In International Conference on Computer Vision (ICCV).

Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. Recycle-GAN: Unsupervised Video Retargeting. In European Conference on Computer Vision (ECCV).

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. CoRR abs/1809.11096 (2018). arXiv:1809.11096 http://arxiv.org/abs/1809.11096

Dan Casas, Marco Volino, John Colliot, and Adrian Hilton. 2014. 4D Video Textures for Interactive Character Appearance. Comput. Graph. Forum 33, 2 (2014), 371–380.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody Dance Now. In International Conference on Computer Vision (ICCV).
Fig. 14. Our results of head-swap. The conditioning image of the head is shown on the bottom for each generated image.

Fig. 15. Failure cases. (Left) Encoding the appearance with a vector makes our method difficult to reconstruct highly entangled spatial features such as text. (Right) Bad conditioning DensePose image during the test time deteriorates the output. Also, see Fig. 16 for challenging cases.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Computer Vision and Pattern Recognition (CVPR).

Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A variational u-net for conditional appearance and shape generation. In Computer Vision and Pattern Recognition (CVPR). 8857–8866.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2020. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. https://arxiv.org/abs/2012.03065 (2020).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In Advances in Neural Information Processing Systems (NeurIPS). 2672–2680.

Artur Grigor’ev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lemptskey. 2019. Coordinate-Based Texture Inpainting for Pose-Guided Human Image Generation. Computer Vision and Pattern Recognition (CVPR) (2019), 12127–12136.

Rieza Alp Gueler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation In The Wild. In Computer Vision and Pattern Recognition (CVPR).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In Conference on Computer Vision and Pattern Recognition (CVPR).

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

Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningsson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus Magnor. 2020. High-Fidelity Neural Human Motion Transfer from Monocular Video. arXiv:2012.10974 [cs.CV]

Tero Karras, Samuli Laine, and Jaakkko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variatiaon. In International Conference on Learning Representations (ICLR).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In Computer Vision and Pattern Recognition (CVPR). 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakkko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In Computer Vision and Pattern Recognition (CVPR).

Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardtz, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. ACM Transactions on Graphics (TOG) 37 (2018).

Diederick P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR).

Bernhard Kratzwald, Zhizhu Huang, Danda Pani Paulel, and Luc Van Gool. 2017. Towards an Understanding of Our World by GANing Videos in the Wild. (2017). https://arxiv.org/abs/1711.11453 arXiv:1711.11453.

Verica Lazova, Eldar Insaftutdinov, and Gerard Pons-Moll. 2019. 360-Degree Textures of People in Clothing from a Single Image. International Conference on 3D Vision (3DV) (2019), 643–653.

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. VOGUE: Try-On by StyleGAN Interpolation Optimization. arXiv preprint arXiv:2101.02285 (2021).

Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2020. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. https://arxiv.org/abs/2011.13084 (2020).

Lingjie Liu, Jiasuo Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020b. Neural sparse voxel fields. In Advances in Neural Information Processing Systems (NeurIPS), Vol. 33.

Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Weipeng Wang, and Christian Theobalt. 2020a. Neural Human Video Rendering by Learning Dynamic Textures and Rendering-to-Video Translation. IEEE
Fig. 16. Highly challenging cases. Even though our method struggles under extreme occlusions, target poses and textures, it synthesizes more plausible images than the compared methods.
Embedding for Semantic Control. ACM Transactions on Graphics (Proceedings SIGGRAPH Asia) 39, 6 (2020).
Justus Thies, Michael Zollhoefer, and Matthias Nießner. 2019. Deferred neural rendering:
image synthesis using neural textures. ACM Transactions on Graphics (TOG) 38 (2019).
Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2020. Non-Rigid Neural Radiance Fields: Recon-
struction and Novel View Synthesis of a Deforming Scene from Monocular Video.
https://arxiv.org/abs/2012.12347 (2020).
Guil Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from Synthetic Humans. In Computer Vision and Pattern Recognition (CVPR).
Marco Volino, Dan Casas, John Collomosse, and Adrian Hilton. 2014. Optimal Rep-
resentation of Multiple View Video. In Proceedings of the British Machine Vision Conference. BMVA Press.
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. Video-to-Video Synthesis. In Advances in Neural Information Processing Systems (NeurIPS).
Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018b. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Computer Vision and Pattern Recognition (CVPR).
Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhöfer. 2020. Learning Compositional Radiance Fields of Dynamic Human Heads. arXiv:2012.09955 [cs.CV].
Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2020. Space-time Neural Irradiance Fields for Free-Viewpoint Video. https://arxiv.org/abs/2011.12950 (2020).
Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. 2011. Video-based Characters: Creating New Human Performances from a Multi-view Video Database. In ACM SIGGRAPH 32:1–32:10.
Zili Yi, Hao Zhang, Ping Tan, and Minglu Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In International Conference on Computer Vision (ICCV). 2868–2876.
Jae Shan Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. 2020. Pose-Guided Human Animation from a Single Image in the Wild. arXiv:2012.03796 [cs.CV].
Polina Zablotskaiia, Aliaksaandr Siarohin, Leonid Sigal, and Bo Zhao. 2019. DwNet: Dense warp-based network for pose-guided human video generation. In British Machine Vision Conference (BMVC).
Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks (Proceedings of Machine Learning Research (PMLR)). 7354–7363.
Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. NERF++: Analyzing and Improving Neural Radiance Fields. https://arxiv.org/abs/2010.07492 (2020).
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Computer Vision and Pattern Recognition (CVPR).
Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612.
Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-
Image Translation using Cycle-Consistent Adversarial Networks. In International Conference on Computer Vision (ICCV).
Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. 2019. Progressive Pose Attention Transfer for Person Image Generation. In Computer Vision and Pattern Recognition (CVPR). 2347–2356.