Human Reposing and Virtual Try-On from Multi-View Images

Jiayun Wang* [1] Amin Kheradmand[2] Himanshu Arora[2]

[1]UC Berkeley [2]Amazon

peterwg@berkeley.edu

Abstract

We study human reposing and virtual try-on from multi-view images. Unlike existing works which take a single image as an input, we learn from multi-view images which are readily available in most fashion datasets and provide rich information on the geometric structures and textures of the human and garments. To this end, since each input view provides partial observation of the person and underlying garments, an appropriate design of parsing and fusing multi-view information for the target image proves essential. We propose a novel framework for warping and fusing reference human images from multiple and varied viewpoints and poses to a target viewpoint and pose. The framework is effective as it considers both 3D human body geometry and 2D photorealism. We also introduce a conditional patch discriminator to further improve image quality. The proposed method outperforms state-of-the-art single-view methods. Specifically, in our experiments with the DeepFashion dataset, we show significant improvements in terms of visual quality, PSNR, SSIM, FID and LPIPS metrics over the existing state-of-the-art approaches.

1. Introduction

In this paper, we study the problem of photo-realistic human image synthesis. The task is becoming increasingly popular due to its wide-range applications, including novel view synthesis, virtual try-on, human reposing, motion transfer and avatar creation, to name a few. For most human image synthesis applications, reference images are needed to guide the target human pose and appearance (including face, skin and garment appearances).

Human image synthesis is a challenging task because: (1) Human body consists of several articulated parts. It is nontrivial to synthesize a human image that accurately represents the 3D structure of individual body parts as well as the holistic relationships among them. In other words, accurately incorporating the geometry and part relationships is important for the realism of a synthesized human image.

*To whom any correspondence should be addressed.

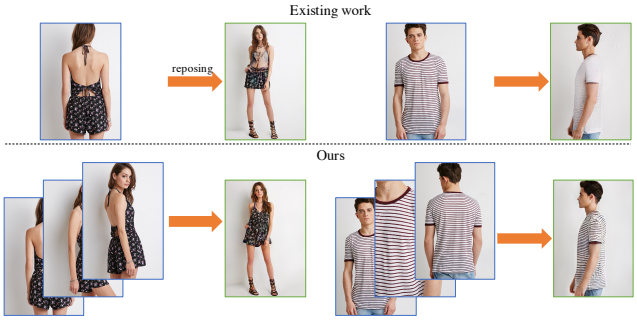

Figure 1. To synthesize a human image with a different pose, existing work [2] takes a single image as input, whereas we show synthesizing from multi-view images leads to higher performance. For instance, the synthesis results benefit from additional observations of the human (left) or the design of our approach that enhances the synthesized garment texture details (right).

(2) Realistic garments are hard to synthesize. As non-rigid objects, the garments may fold and their texture may alter differently on different parts of the human body. This greatly increases the difficulty of rendering human images with high fidelity and photorealism.

Unlike existing methods which take a single-view reference image as input, our approach alleviates the aforementioned challenges by learning from multi-view reference images (Fig.1). Synthesizing images from multi-view inputs has several advantages: (1) Multi-view human images are common, easy to obtain and readily available. For example, in the fashion industry, it is common to shoot a certain garment from multiple views for a better visualization experience. This provides large-scale readily-available images for designing multi-view human image synthesis systems. (2) The geometric and holistic structures of a human body can be better parsed, represented and synthesized from multi-view input images. (3) The whole garment is never visible from a single view. Also, for many poses, different body parts occlude garments. Using multiple views provides additional observations which enable the reconstruction of the entire texture of the garment in high fidelity.

Synthesizing novel images from multi-view inputs has shown great potential and success for objects and scenes

[4,25,33], but has not been widely adopted in human image synthesis. Existing human image synthesis works mainly focus on synthesizing from a single image [2, 11]. They are able to successfully synthesize human images when the view change from the reference to the target is not drastic. [38] further improves the synthesis quality with garment warping techniques. However, the quality suffers with drastic view changes due to occlusions and limited visible regions from a single view. We propose a novel approach that efficiently parses and fuses multiple input images for synthesis. Some recent works reconstruct 3D models from multi-view person images [31, 39] and use these models to render human images from novel viewpoints. These approaches provide high-quality human body shape reconstruction. However, they are not able to generate the texture maps accurately due to the difficulty of registration among different views. In this paper, we directly synthesize human images to enhance textural quality and photorealism.

Our method takes a pose image (pose reference) and multi-view human images (appearance reference) as inputs[1]. Firstly, we warp multi-view images to the target pose and bring all images to the same pose. We perform the warping in both 3D human body UV space and image pixel space to take advantage of the respective domains. We base our UV space warping on DensePose [12], a popular 3D geometry-aware human pose estimation algorithm. For pixel-space warping, we use the thin-plate spline (TPS) transformation [7], which takes into account correspondences in the pixel domain to deform the garments from the source poses to the target pose. While using UV space warping from multiple views allows texture reconstruction of all parts of the target pose including skin pixels, it usually fails to preserve texture details. Pixel space warping is not always complete, however, it preserves the texture details since it uses an image domain transformation (TPS). We encode the warped multi-view human images as appearance features, and fuse them using visibility confidence maps. In our experiments, we show that the proposed design of multi-view fusion is essential for the success of the synthesis. Finally, we decode the encoded appearance and pose features to generate the synthesized image. To preserve the realism of the generated images while preserving the generated garment texture, we propose using a patch-based discriminator that uses patches from both appearance sources and the generated image and encourages their co-occurrence.

We summarize our contributions as follows: (1) We design a novel multi-view fusion approach that is non-trivial and necessary for a successful multi-view human image synthesis. Our results demonstrate that multi-view human image synthesis significantly outperforms single-view synthesis. To the best of our knowledge, we are the first to study multi-view human reposing and virtual try-on. (2) In contrast to previous methods that either use warping in UV space or pixel space, we propose to warp source images in both UV and pixel space to take advantage of the benefits of both approaches to improve the rendering quality of the human skin as well as the garment textures. (3) We introduce a conditional patch loss that improves the fidelity and details of the generated images.

[1]Our approach also works for a single-view image as the input, which can be considered as a special case of the multi-view inputs.

2. Related Work

Image-Based Clothed Human Synthesis. Clothed human image synthesis has recently attracted much attention. With the success of StyleGAN and its variants for generating high-quality images [17, 18], most state-of-the-art human image synthesis methods [8, 32] adopt StyleGAN as the backbone network or inherit it as their fundamental building blocks. In addition to unconditional human image synthesis methods aforementioned, conditional synthesis methods have also been developed. Such human image synthesis methods are usually conditioned on human poses, enabling applications like human reposing [2, 23, 24], motion transfer [1,41] and virtual try-on [32,38]. Specifically, PoseGAN [20] adopts StyleGAN architecture and conditions it on estimated human pose images. Pose-with-Style (PWS) [2] builds upon PoseGAN and further warps source appearance images to the target pose with a pose-guided appearance flow. They show improved performance when adding the warped appearance images to StyleGAN. All these existing works take a single image as input. In this paper, we explore using multi-view images as input as they provide more information for improving the quality of the synthesis task.

Novel View Synthesis from Multi-View Images. Novel view image synthesis has been a well-studied task in the literature. Tatarchenko et al. [35] directly generate the target-view image, while Zhou et al. [44] consider the task of novel view synthesis as predicting dense flow fields that map the pixels from the source view to the target view. Neural rendering [25, 42] represents another promising direction. [22, 25] learn volumetric neural scene representations for novel view synthesis. HumanNeRF [37] extends the idea to human image synthesis and generates novel human images from monocular human videos. Neural rendering methods usually need to be trained on each specific example, and thus are not time-efficient for inference. Other researchers also study direct reconstruction of 3D shapes from source images, with more details in the next subsection.

Human image synthesis from multi-view images has not been actively studied before. To the best of our knowledge, we are the first to study this task.

3D Human Reconstruction. Another category of methods directly reconstruct a 3D model (e.g., mesh with texture) from a single human image. Novel-view images then can be rendered from the 3D model. SfM [26] and SLAM [9]

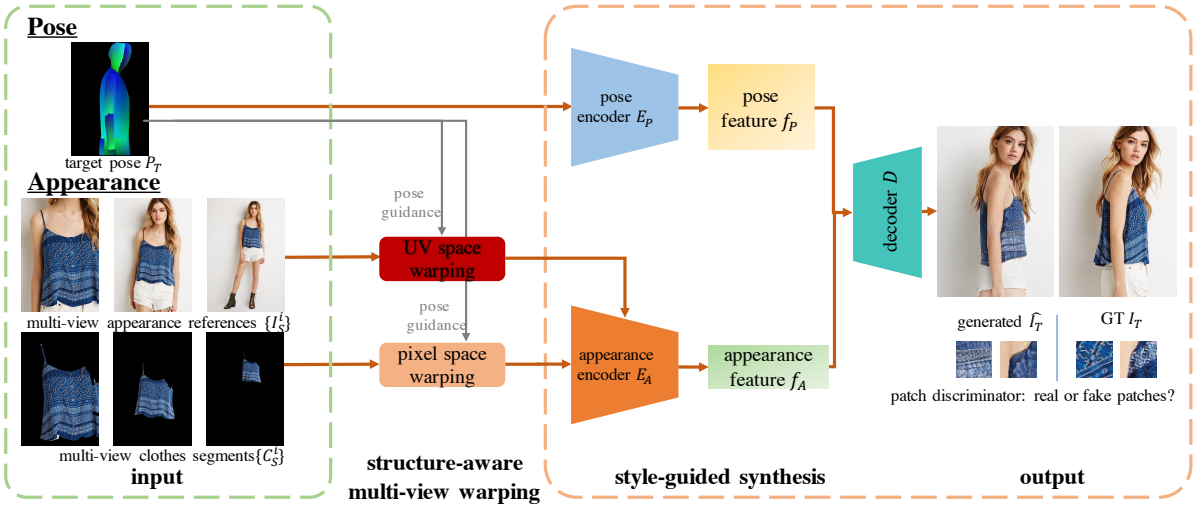

Figure 2. Overview of our method for human reposing. We take a pose image P_T as the reference for the target human pose, as well as multi-view human images $\{I_S^i\}$ as the reference for the target human appearance, i.e., the face identity, the skin tone and the garment texture. Pose reference P_T is fed to the pose encoder E_P for the pose feature f_P. Appearance reference images $\{I_S^i\}$ and their corresponding garment segments $\{C_S^i\}$ are warped to the target pose in both UV and pixel space (Section 3.1), and encoded as an appearance features $\{f_A^i\}$ with the appearance encoder E_A. $\{f_A^i\}$ of every view is fused to an aggregated appearance feature f^A based on visibility from each source to the target pose (Section 3.2). The StyleGAN2-based [18] decoder D takes the structure and the appearance feature f_P, f_A and decodes them as the output image $\hat{I}_T$. To improve the synthesis quality, we supplement a conditional patch discriminator to the generative adversarial networks (Section 3.3). The method could be easily adapted for virtual try-on during inference.

successfully handle multi-view 3D reconstructions in various real-world scenarios, but their performance may suffer when the input data is scarce or is not representative of varied viewpoints. Recently, deep learning methods have taken leads in further improving the reconstruction quality by completing the occluded or hollowed-out areas [16, 40]. Specifically, for 3D human reconstruction, SMPL [5] reconstructs 3D human vertex-based model which is able to handle a wide variety of body shapes in natural human poses. PiFU [31] further estimates the human body texture from the image, and is able to handle single or multi-view inputs. While researchers are working to further improve the reconstruction quality [3, 39], the texture of the reconstructed model is often not of high quality. In contrast, we do not use the 3D model reconstruction as an intermediate step to synthesize the human image. Instead, we directly synthesize novel-view images from multi-view inputs.

Improving the Quality of Synthesized Garments. Efforts have been made to further improve garment quality in the synthesized images. VITON [13] first integrates a UNet-based [30] generation network to its deformation-based approach for garment synthesis in a novel view. Wang et al. [36] first deform the source garment to the target shape and then synthesize the try-on result. PASTA-GAN [38] focuses on the frontal view for human image synthesis, and proposes a patch-based method for garment deformation. On the contrary, we propose a new garment deformation method, which works for drastic view changes as well and is capable of taking multi-view garment images as input.

3. Methods

We illustrate the overall human image synthesis framework in Fig.2. Two references are fed to the network for image synthesis: target pose P_T and multi-view appearance images $\{I_S^i\}$. Multi-view appearance references are warped to the target pose, encoded and fused to the appearance feature f_A. The target pose is encoded as the pose feature f_P. A decoder takes both f_A and f_P to decode the output image.

We introduce each component of the framework as follows: we first discuss how multi-view appearance reference images are warped to match the target pose in Section 3.1, and then present the encoder that fuses multiple views and the decoder in Section 3.2. Finally, in Section 3.3, we provide details on the training procedure and the loss functions.

3.1. Source-to-Target Warping

Let us consider a special case where only a single source image I_S exists. The appearance reference image I_S and the target image I_T to be synthesized usually have different human poses which we denote by P_S and P_T, respectively. To improve the quality and realism of the synthesized image, we warp the source image I_S and garments C_S, so that their human pose matches with the target pose (Fig.3). We present warping details in Sections 3.1.1 and 3.1.2.

Now we consider the case where N multi-view source images $\{I_S^i\}_{i=1,2,...,N}$ are used as inputs. For the i^{th} view, we estimate the human pose P_S^i with DensePose [12] from the source appearance image I_S^i. Firstly, based on the correspondence between the source and target poses P_S^i, P_T, we

Figure 3. Appearance warping in UV and pixel spaces. Multi-view appearance reference images $\{I_S^i\}$ and corresponding garment segments $\{C_S^i\}$ are warped from the source pose to the target pose for improving the synthesis quality. Specifically, image I_S^i is warped to target pose $W(I_S^i)$ in the UV space based on the inpainted human texture UV map. Garment C_S^i is warped in the pixel space to the target pose $T(C_S^i)$. The warping in both spaces guarantees the synthesis quality in terms of both 3D human body geometry and 2D photorealism. Based on the common regions between the source and the target pose, the visibility map V_S^i is also generated to aid the multi-view fusion.

compute the visibility map V^i, which indicates the regions in the target image that is also available in the i^{th} source image. Additionally, as discussed before, the source image I_S^i is warped as $W(I_S^i)$ to match the target pose, and the source garment C_S^i is also warped as $T(C_S^i)$. All N images are encoded as features and fused for image synthesis, which we discuss in detail in Section 3.2.

3.1.1 Warping in UV Space

We warp input appearance image I_S^i of the i^{th} view to match the target pose P_T. The UV map of a human pose P_S^i from the source view i makes it possible to obtain pose-agnostic 3D-human-body-shape-aware UV texture map. One issue here is that the texture map is missing the regions that are not available in I_S^i due to occlusion. We thus follow [2] to inpaint the missing regions and obtain the full human texture map T_S^i. Further, we introduce a *confidence map* to help the network better understand which regions in the warped image are inpainted and which regions are derived directly from the source image. Specifically, based on the correspondence between the source pose P_S^i and the target

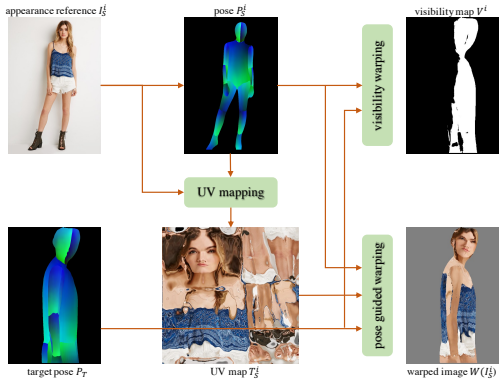

Figure 4. 3D-aware warping in UV space. The goal is to warp i^{th} source image I_S^i to the target pose P_T. First, based on the correspondence between the estimated source pose P_S^i and target pose P_T, the visibility map V_i is computed. Then, based on the source pose P_S^i, the source image I_S^i is mapped to the UV-space, and inpainted to obtain the UV texture map T_S^i. UV map can be mapped to 2D $W(I_S^i)$, which is the warped image. $W(I_S^i)$ and V_i are later used for multi-view fusion.

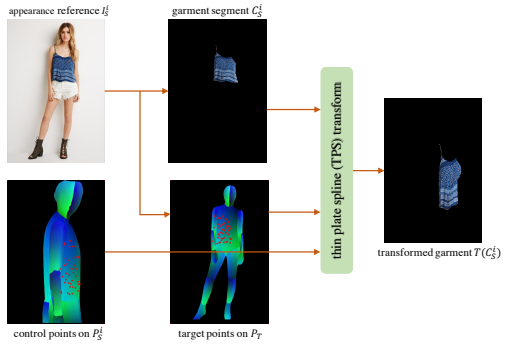

Figure 5. 2D garment spatial deformation. The garment C_S^i is segmented from i^{th} view source image I_S^i. To warp C_S^i so the garment fits the target pose, corresponding control points Q_S^i and target points Q_T^i are sampled from the source pose P_S^i and the target pose T_i^S, respectively. Guided by the point pairs, the source garment C_S^i is then warped to the target pose with thin plate spline (TPS) transform [7] . Note that the upper body and the lower body garment (not depicted in the figure) are segmented and warped separately, and are both fed to the network for synthesis.

pose P_T, we compute the visibility map V_i, which will be later used in our multi-view fusion (Section 3.2). Since we also use the target pose P_T as input, corresponding pixels can then be mapped to the target pose from the texture map T_S^i to obtain the warped image $W(I_S^i)$. The warping procedures are also illustrated in Fig.4.

3.1.2 Warping in Pixel Space

We warp garments from the i^{th} source pose P_S^i to the target pose P_T to provide additional guidance for the network to synthesize the target garments. PASTA-GAN [38] is one of the first works which demonstrates that adding additional deformed garments improves the quality of the synthesized

clothed human image. However, PASTA-GAN cannot handle drastic pose changes (e.g., front view to side view) because: (1) Their patch-based garment deformation is designed for the full-body frontal view of a human. It is hard to generalize to difficult cases where drastic view changes are present. (2) Their deformation is based on OpenPose [6] which encodes body skeletal pose and hence cannot differentiate front vs. back or different 3D body shapes.

We propose a DensePose-based garment deformation method which works well with drastic view changes (Fig.5). Our method can handle human poses from any viewpoint and is aware of the 3D shape of the human body.

Garment Segmentation. For a source appearance image I_S^i of the i^{th} view, we first segment the upper body garment and the lower body garment with Graphonomy [10], which is an off-the-shelf clothed human image segmentation algorithm. The upper and lower body garments go through the same procedure for deforming to the target pose. Here, we only discuss one of the segmented garments denoted by C_S^i.

Obtaining Corresponding Points from Poses. To guide the garment deformation, we find corresponding points on the source pose P_S^i and the target pose P_T. For this we find grid points in the UV space of source and target poses P_S^i, P_T^i [2]. Each point in the grid has two values for $\{u, v\}$ and the corresponding pixel coordinate values $\{x, y\}$. Grid points on P_T, P_S that share the same UV values refer to the same locations in the 3D human body defined by DensePose [12]. Due to occlusion, only visible grid points are mapped to 2D pixel space. J points that are available in both source and target images are thus found and considered as correspondences. We denote such coordinates on source image as control points $Q_S^i = \{x_S^{ij}, y_S^{ij}\}_{j=1,...,J}$, and corresponding coordinates on target image as target points $Q_T^i = \{x_T^{ij}, y_T^{ij}\}_{j=1,...,J}$.

Garment Deformation. If there is a drastic change between the source and the target pose (e.g. front and back), we do not deform garments and only use UV warping results to guide the synthesis. When the number of control points J is larger than 30 (a heuristic threshold we adopt in the paper), we use thin plate spline (TPS) transformation [7] for garment deformation. Specifically, TPS transforms source garment C_S^i to the target pose using the source control points Q_S and corresponding target points Q_T. Examples of the deformed garment $T(C_S^i)$ are in Fig.5.

3.2. Multi-View Fusion and Image Synthesis

Encoding Images to Features. We take N images $\{I_S^i\}_{i=1,...,N}$ as inputs. After warping, we obtain warped source images $\{W(I_S^i)\}$ and garments $\{T(C_S^i)\}$ [3], which are then encoded to multi-scale features following StyleGAN2 [18]. Specifically, at layer l, the pose feature ${}^l f_P$ is a convolutional feature of target pose P_T, while the appearance feature of the i^{th} view ${}^l f_P^i$ is a concatenated feature of the warped image $W(I_S^i)$ and garment $T(C_S^i)$.

Multi-View Fusion. We fuse all appearance features $\{{}^l f_A^i\}$ from N input views by weighted averaging. The visibility map V^i is a binary mask indicating the common regions between the i^{th} source pose and the target pose. We use two 3×3 convolutions separated by a ReLU activation function to transform the concatenated visibility map V^i and warped image $W(I_S^i)$ to a learned single-channel confidence map U^i. U^i is then normalized to $U^{i\prime} = \{u_{x,y}^{i\prime}\}$, such that the sum of the N views of each pixel location $u_{x,y}^{i\prime}$ is equal to 1. That is, $\sum_i^N u_{x,y}^{i\prime} = 1$. At each layer l, $U^{i\prime}$ is resized to ${}^l U^{i\prime}$, which has the same spatial size as the appearance feature map ${}^l f_A$. The fused appearance feature ${}^l f_A$ is then a sum of the products of each of the N view-dependent appearance feature maps ${}^l f_A^i$ with their corresponding weight ${}^l U^{i\prime}$ at each pixel location: ${}^l f_A = \sum_i^N {}^l U^{i\prime} \circ {}^l f_A^i$, where $\circ$ refers to the element-wise product for matrices.

Decoding to the Target Image. We use StyleGAN2 [18] blocks with spatial modulation to decode multi-scale features to images. We use a similar spatial modulation as in [2] to preserve spatial structures of appearance feature f_A. Specifically, at layer l, the scaling α and bias β are generated with 1×1 convolutions from appearance features ${}^l f_A$. Pose feature is then modulated as ${}^l f'_P = \alpha\, {}^l f_P + \beta$, and normalized as ${}^l f''_P = \frac{{}^l f'_P - \text{mean}({}^l f'_P)}{\text{std}({}^l f'_P)}$.

3.3. Patch Discriminator and Loss Functions

Conditional Patch Discriminator. Unconditional patch discriminator [28] enforces co-occurrent patch statistics across different regions of the image, and demonstrates improved textural quality in image synthesis. We improve upon it and adopt a conditional patch discriminator D_{patch} in addition to the full image discriminator. The goal is to enforce the realism of patches as well as being similarity of patches in the reference image. Given image patches of a source image patch(I_S) (whose view is randomly selected), the discriminator is asked to distinguish between image patches from the generated image patch$(\hat{I_T})$ and those from the ground truth I_T. The corresponding patch loss for the discriminator is calculated as:

$$L_{\text{patch}} = \mathbb{E}[-\log(D_{\text{patch}}(\text{patch}(\hat{I_T}), \text{patch}(I_T)|\text{patch}(I_S)))] \quad (1)$$

where D_{patch} refers to KL-divergence that measures the distance between the generated and ground-truth patches. Empirically, we show improved performance with conditional patch loss compared to unconditional patch loss $\mathbb{E}[-\log(D_{\text{patch}}(\text{patch}(\hat{I_T}), \text{patch}(I_T)))]$ (Section 4.6).

[2] The sampling interval of the x and y grid locations is 10 in a 256×256 UV map.

[3] As discussed before, we process both upper-body garment and lower-body garments. For convenience, we denote the concatenated version of the two warped garments as $\{T(C_S^i)\}$.

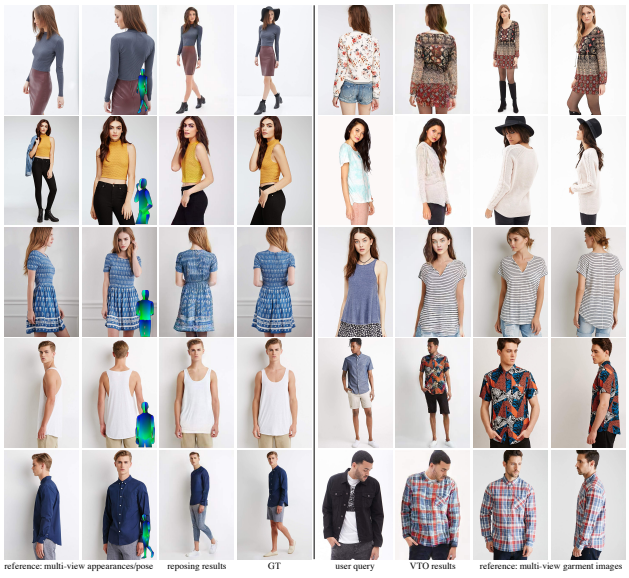

Figure 6. Visual results on human reposing (left) and virtual try-on (VTO, right). Our method synthesizes human images with high fidelity in terms of geometric structures and garment textures.

Training Losses. In addition to the patch loss L_{patch}, the full image adversarial loss L_{GAN} used in StyleGAN2 [18], and the face identity loss L_{face} used in PWS [2], we also minimize the $\mathcal{L}_1$ distance between the generated image $\hat{I_T}$ and ground-truth I_T in both pixel and feature space:

$$L_{\text{rec}} = \|\hat{I_T} - I_T\|_1 + \sum_{l=1}^{k} \|\phi_l(\hat{I_T}) - \phi_l(I_T)\|_1 \quad (2)$$

where $k = 5$, ϕ_l denotes the l^{th} feature map in a VGG-19 [34] pretrained on ImageNet. Here, we use 5 feature maps. Therefore, L_{rec} minimizes the differences between the generated image and the ground truth in terms of both pixel appearance and pre-trained features.

Our final training loss is a combination of the above-mentioned loss functions: $L = L_{\text{GAN}} + \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{face}} + \lambda_3 L_{\text{patch}}$. For experiments, we set $\lambda_1 = 5, \lambda_2 = 1, \lambda_3 = 1$.

4. Experimental Results

4.1. Implementation and Training Details

Implementation details. For human reposing, source images $\{I_S^i\}$ with spatial size 512×512 were first fed to a segmentation network [10] to obtain garment images $\{C_S^i\}$, which are then deformed with TPS [7]. Each source image I_S^i and deformed garment image $T(C_S^i)$ were fed to an appearance encoder E_A consisting of one convolution layer and 5 ResNet [14] blocks with numbers of channels $64, 128, 256, 512, 512, 512$ and output spatial sizes $512, 256, 128, 64, 32, 16$. At each block, feature maps of I_S^i are warped in the UV space to match the target pose P_T. For every viewpoint, appearance features of both image I_S^i and garment C_S^i are concatenated and fused following procedures mentioned in Section 3.2. As another input, target pose P_T is fed to a pose encoder E_P, with the same architecture as E_A. To decode to the target image, at each block, pose and structure features are fed to a StyleGAN2 [18] block with spatial modulation mentioned in Section 3.2. We adopt the same decoder architecture as [2].

For virtual try-on, instead of a target pose P_T, a query image I_Q is fed to the network. Target pose is estimated from I_Q, and fed to the pose encoder E_P. Also, I_Q is fed to the appearance encoder E_A, together with garment reference $\{I_S^i\}$ and $\{C_S^i\}$. A garment mask M estimated from I_Q controls the regions to apply garment references. Our experiments focus on virtual try-on for the whole body, though the framework can generalize to partial body try-on.

Training Details. We use Adam optimizer [19] with a learning rate of $\eta \cdot 0.002$ and $\beta = (0, 0.99^{\eta})$. For the generator, $\eta = 0.8$. For the full image and patch discriminator, $\eta = 0.9$. The model is implemented with PyTorch framework. The batch size is 8. We first train the model with loss functions only on the human region of the output image for 50 epochs. We then fine-tune the model with loss functions on the entire image for another 10 epochs. In total, our training took 7 days on 8 NVIDIA V100 GPUs.

Inference Time. On a NVIDIA V100 GPU, generating a reposing image takes on average 1.4 seconds while the virtual try-on takes 1.9 seconds.

4.2. Datasets and Evaluation

Datasets. DeepFashion dataset [21] is used in this paper for training and evaluation. We follow the same training and evaluation split of PWS [2] for reposing. For virtual try-on, we choose random pairs from the evaluation set. Specifically, there are 101,967 training, 8,570 evaluation pairs for reposing and 8,570 evaluation pairs for virtual try-on.

Baselines. We compare with state-of-the-art methods [2, 24, 27, 29, 38]. For fairness, we follow the respective settings of the above methods, train and evaluate on the same split of the DeepFashion dataset [21]. We also make a stronger variant of PWS [2] that we call as PWS-closest pose. It takes all available multi-view images, uses DensePose [12] to determine the most similar input pose to the target pose, and uses the input image with the closest pose for synthesis.

Evaluation Protocols. Following [2], we use the human foreground peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS) [43], and Frechet Inception Distance (FID) [15]. We report these metrics for comprehensiveness and fairness when comparing with other methods.

4.3. Reposing

We report quantitative results in Table 1 and some visual examples in Fig.6 and Fig.7. Our method compares fa-

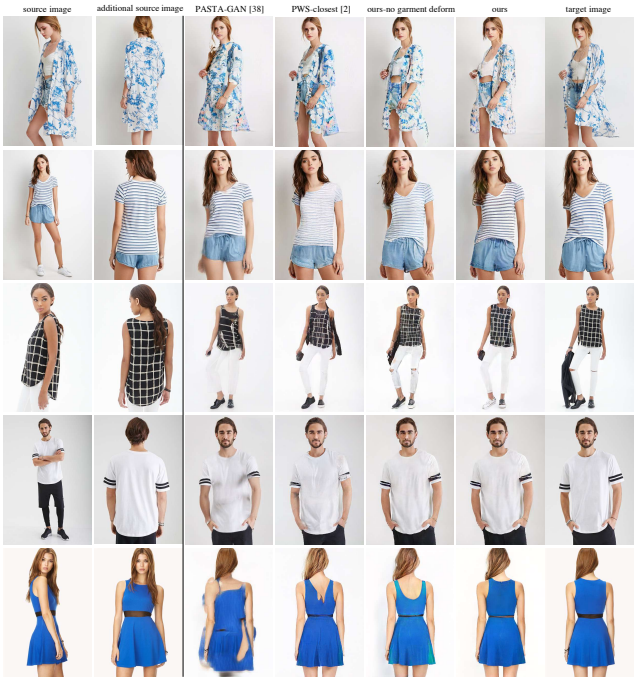

Figure 7. Human reposing comparisons. For fairness, PWS [2] and PASTA-GAN [38] consider all available input images and takes the one with the closest pose (1st column, source image). Our method takes multi-view images as input (with a sample additional view in column 2) and outperforms baseline methods. We synthesize images with higher fidelity, better geometric structures and more realistic garment textures.

Table 1. Human reposing image synthesis results on DeepFashion dataset [21]. Our method outperforms SOTA methods.

	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
PASTA-GAN [38]	14.51	0.49	34.54	0.170
PATN [27]	17.70	0.75	21.86	0.195
ADGAN [24]	17.72	0.75	16.27	0.175
GFLA [29]	18.04	0.76	15.17	0.167
PWS [2]	18.50	0.77	9.40	0.134
PWS [2]-closest pose	18.92	0.78	9.07	0.096
Ours (MV + dual warp.)	**19.77**	**0.82**	**8.37**	**0.083**

vorably against existing works such as PASTA-GAN [38] and PWS [2]. Note that ours also outperforms a multi-view baseline method, PWS-closest pose, which considers all available input views and synthesizes from the one with the closest pose. We observe improved visual quality of the synthesized images with our method, especially in terms of geometry and garment structure details.

Table 2. Our method outperforms SOTA for virtual try-on.

	PASTA-GAN [38]	PWS [2]	PWS closest pose	ours
FID	34.64	26.53	21.11	20.08

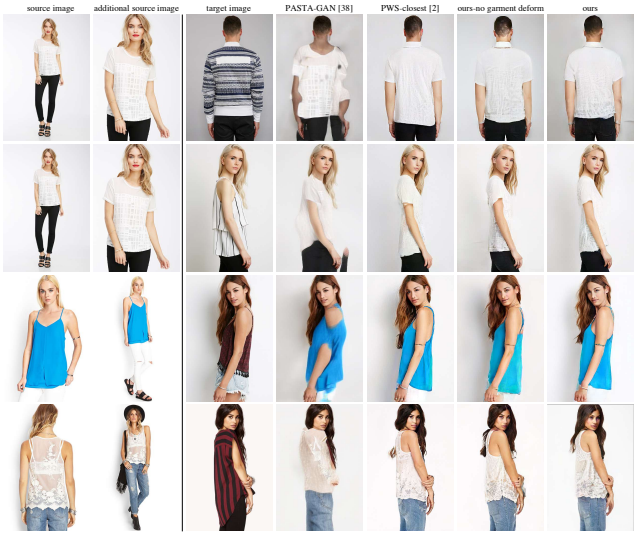

Figure 8. Virtual try-on comparisons. All methods consider all available input views. Our method outperforms baseline methods with synthesized images of higher fidelity.

4.4. Virtual Try-On

We report quantitative comparisons with baseline methods in Table 2. Our method outperforms baseline methods [2, 38]. For visual comparison results in Fig.8, our synthesized images have better visual quality, especially higher fidelity in the geometric details and textures.

4.5. Synthesizing with Various Views and Poses

The proposed model allows us to re-render a human image with a variety of viewpoints and many poses (Fig.9), thanks to the multi-view inputs. Given multi-view source images, we fit a 3D human model using PIFu [31]. When rendered from novel viewpoints, the generated texture for this model is not of high quality and not realistic. To generate realistic images from novel viewpoints, we first estimate DensePose [12] P_T from the PIFu rendered images, and use our algorithm to synthesize human images. Fig.9 (left) shows such example synthesized images. Note that the quality may be limited by the 3D human reconstruction performance, for instance, the left hand in Fig.9 is not well reconstructed by PIFu [31], and leads to lower synthesized image quality in the hand region. For different poses, we estimate the target pose P_T from another image in the DeepFashion [21] evaluation set, which has a different human pose (see Fig.9 right).

4.6. Ablation Study and Analysis

We conduct ablation studies and different design variant analysis on the novel parts of the proposed method.

Garment Warping Module. We consider three variants: (1) No garment warping, (2) OpenPose-based [6] garment warping, and (3) The proposed garment warping with TPS

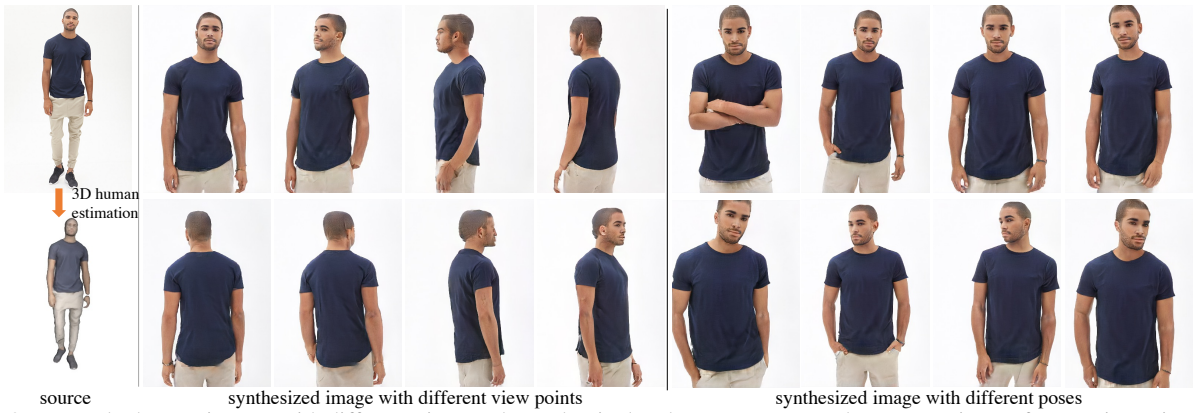

Figure 9. Re-render human images with different view angles and articulated poses. We re-render a source image from various view angles (by fitting a 3D human body model [31], lower left). We also re-render the source image with different articulated poses.

Table 3. Ablation studies and analyses of design variants of (a) the garment warping module, (b) the multi-view fusion and (c) patch discriminator.

(a) Garment Warping

	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
no garment warp.	19.03	0.79	8.81	0.091
OpenPose [6] garment warp.	19.34	0.81	8.67	0.089
our TPS garment warp.	**19.77**	**0.82**	**8.37**	**0.083**

(b) Multi-View Fusion

	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
no multi-view	18.92	0.78	9.07	0.096
naïve average	19.18	0.80	8.81	0.088
sequential comb.	19.39	0.80	8.52	0.087
our multi-view fusion	**19.77**	**0.82**	**8.37**	**0.083**

(c) Patch Discriminator

	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
no patch	19.59	0.81	8.71	0.086
uncond. patch	19.60	0.81	8.70	0.086
our cond. patch	**19.77**	**0.82**	**8.37**	**0.083**

[7]. Specifically, for (2), we adopt a similar patch-based garment warping method with OpenPose serving as the correspondence between the source and target image. According to the quantitative comparison (Table 3(a)) and the visual comparison (Fig.7 and Fig.8), the proposed DensePose-based garment warping with TPS transformation has the best performance. We hypothesize that the performance gain comes from the 3D-human-geometry-aware DensePose estimation, which is especially useful for source-to-target garment warping when the view change is drastic.

Multi-View Fusion Variants. We consider four variants: (1) No multi-view inputs. The model just takes a single input source image which has the closest pose to the target pose in terms of DensePose [12] estimation. (2) Naïve average. For the appearance feature f_A^i of each input view, the model simply averages all features. (3) Sequential combination. For each view, we compute the visibility map V^i, which indicates the common region that is available in both source and target image. To fuse $\{f_A^i\}$, we first rank all N views from the closest to the furthest to the target image. The fused appearance feature f_A is a sequential combination of $f_A^1 \circ V^1, ..., f_A^N \circ V^N$, where $1, ..., N$ is ranked from the closest to the furthest view. Finally, for regions that are still not covered (i.e. not available in any input view), we supplement with corresponding (inpainted) regions of the closest view feature f_A^1. (4) Our proposed confidence-map-based multi-view appearance feature f_A^i fusion method (details in Section 3.2). As reported in Table 3(b), the proposed confidence-map-based feature fusion mechanism outperforms other variants. We hypothesize the reason is that the mechanism considers the common regions between each source and target image.

Conditional Patch Discriminator. We consider three variants: (1) No patch discriminator. (2) Unconditional patch discriminator with loss $L_{\text{patch}} = \mathbb{E}[-\log(D_{\text{patch}}(\text{patch}(\hat{I_T}), \text{patch}(I_T))]$. (3) Our proposed conditional patch discriminator with loss $L_{\text{patch}} = \mathbb{E}[-\log(D_{\text{patch}}(\text{patch}(\hat{I_T}), \text{patch}(I_T))|\text{patch}(I_S)]$. As reported in Table 3(c), the proposed conditional patch discriminator outperforms other variants. Similar to conditional generative adversarial networks, considering the input I_S leads to better image generation quality.

Summary. We present a novel method for human reposing and virtual try-on from multi-view images. The major novelty lies in the multi-view fusion mechanism and the source-to-target warping mechanism in both UV and pixel space. We also introduce a conditional patch-based discriminator, which could be used for generative adversarial networks in general. Experiments on the large-scale DeepFashion dataset show that synthesizing from multi-view images leads to higher fidelity and better geometric details as compared to the single-image approaches. Additionally, the proposed method significantly outperforms state-of-the-art methods, both visually and quantitatively.

For limitations, our method does not always synthesize complicated textures well when all source poses are far from the target. The virtual try-on is also a direct inference on the model trained on reposing only. Additional fine-tuning for virtual try-on could lead to higher visual quality.

References

[1] Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019. 2

[2] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11, 2021. 1, 2, 4, 5, 6, 7

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 3

[4] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5960–5969, 2020. 2

[5] 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*, pages 561–578. Springer, 2016. 3

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5, 7, 8

[7] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977. 2, 4, 5, 6, 8

[8] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7723–7732, 2022. 2

[9] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial intelligence review*, 43(1):55–81, 2015. 2

[10] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 5, 6

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

[12] 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*, pages 7297–7306, 2018. 2, 3, 5, 6, 7, 8

[13] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] 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*, 2017. 6

[16] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, 2017. 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3, 5, 6

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Kanglin Liu, Qing Li, and Guoping Qiu. Posegan: A pose-to-image translation framework for camera localization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:308–315, 2020. 2

[21] 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*, pages 1096–1104, 2016. 6, 7

[22] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[23] 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*, pages 99–108, 2018. 2

[24] 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/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 2, 6, 7

[25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[26] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017. 2

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 6, 7

[28] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 5

[29] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 6, 7

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[31] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3, 7, 8

[32] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

[33] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[35] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *European Conference on Computer Vision*, pages 322–337. Springer, 2016. 2

[36] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 3

[37] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2

[38] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. *Advances in Neural Information Processing Systems*, 34:2598–2610, 2021. 2, 3, 4, 6, 7

[39] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 3

[40] Bo Yang, Stefano Rosa, Andrew Markham, Niki Trigoni, and Hongkai Wen. Dense 3d object reconstruction from a single depth view. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2820–2834, 2018. 3

[41] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 2

[42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[43] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 6

[44] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 2