Poxture: Human Posture Imitation Using Neural Texture

Chen Yang, Shun-Yu Yao, Zan-Wei Zhou, Bin Ji, Guang-Tao Zhai, *Senior Member, IEEE*, and Wei Shen

***Abstract*— Human pose imitation, which aims to generate an image with a source character's appearance, the source character's shape, and a target character's posture, has many potential applications in virtual reality, augmented reality, games, movies, etc. It is incredibly challenging due to non-rigid human body motions, significant variations in clothing textures, and self-occluded human bodies in 2D images. In this paper, we propose Poxture, a novel human posture imitation method with neural texture, to address the challenges mentioned above. Concretely, first, we build a dense mapping between a source SMPL human body model (shape and posture) and its corresponding texture (appearance). Then, we apply a neural texture generator to recover the complete texture of the source character. At last, we wrap the source neural texture to the source SMLP model with a target pose to generate the desired image by a GAN model. Poxture does not require any annotations, and our framework can fully disentangle the source character's appearance, shape, and pose, which enjoys several advantages: 1) It can synthesize high-resolution images with detailed textures, thanks to the learned neural textures containing both visible and invisible parts and high-frequency information; 2) It can imitate complex actions with various appearances and body figures since the complete texture of the source character is acquired. We compare our method with previous methods, showing state-of-the-art results on two challenging benchmarks. Extensive experiments demonstrate that, given any character, our method can manipulate this avatar imitating arbitrary posture.**

***Index Terms*— Motion transfer, human pose imitation, person image generation.**

I. Introduction

HUMAN posture imitation or human pose reenactment aims at combining one person's appearance and shape with another one's posture. Given a source character, human posture imitation aims to manipulate this avatar to present any pose with a target human's image as reference, significantly benefiting downstream applications such as games, movies, virtual or augmented reality. Human posture imitation is motivated by the demand for synthesizing scenes from the film industry. In the past few decades, this task often relied on stunt performers or experts' manual adjustment.

Human posture imitation is exceptionally challenging, as it requires generating precisely the same pose as the reference human and the corresponding detailed texture and shape from the source human simultaneously. Consequently, it is hard to build an end-to-end pipeline to handle this task consummately. Traditional computer graphics based motion transfer methods [1], [2] require complex rendering operations to generate pseudo texture, which are time-consuming and calculation intensive. Fortunately, with the development of machine learning and computer graphics, it is feasible to reconstruct a high-precision 3D human model from a monocular RGB camera. These impressive methods provide radically different solutions for human posture imitation.

Recently, there springs up a large amount of related literature which has acquired convincing results and stunning performance on this challenging task. These methods can be mainly divided into two categories: 2D image warping based methods [3]–[11] and 3D spatial transformation based methods [12]–[15].

2D image warping based methods tackle human pose imitation by learning general warping functions from images with source characters to images with desired poses [3]–[11]. The warping function can transfer human body pixels of the source image to desired coordinates corresponding to the target pose. It is usually learned from 2D clues such as 2D joints [3], [5], [6], 2D skeletons [8], [11], human part segmentation [9], [10], [16] or some unsupervised keypoints [4]. Unfortunately, they generate distorted and unnatural images, especially under significant pose variations, since such a warping function is only able to learn 2D affine transformation. However, human poses and appearances are essentially three-dimensional. Some recent methods [7] apply the warping function on feature spaces to improve the quality of generated images. However, since they cannot avoid the inherent flaws of the warping function, they cannot perform precise imitation.

3D spatial transformation based methods tackle this task in 3D space with different representations of human bodies [12]–[15]. Some 3D transformation methods [14] detect sparse human 3D joints and leverage joints for guiding imitation. However, only with joints, they fail to capture the shapes of human bodies. Other 3D spatial transformation-based methods address this imitation task by virtue of reconstruction and projection. First, they reconstruct the source character's human body model with the help of a dense pose [13], [15] or a parametric human model [12]. Then, they

Manuscript received 29 January 2022; revised 7 May 2022; accepted 30 June 2022. Date of publication 14 July 2022; date of current version 6 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62176159, in part by the Natural Science Foundation of Shanghai 21ZR1432200, and in part by the Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102. This article was recommended by Associate Editor J. Liu. *(Chen Yang and Shun-Yu Yao are co-first authors.) (Corresponding author: Wei Shen.)*

The authors are with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ycyangchen@sjtu.edu.cn; ysy2017@sjtu.edu.cn; sjtu19zzw@sjtu.edu.cn; bin.ji@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn; wei.shen@sjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3190875.

Digital Object Identifier 10.1109/TCSVT.2022.3190875

somehow link human body pixels of the source image and the human body model. Finally, they repose the body model to the target pose and project it to get the result. Since single images contain limited shape and appearance information, the link between pixels and the body model is not authentic. These methods fail to generate the complete texture of one human as well as some physical attributes of the appearance, such as albedo or reflectance. Consequently, they are only able to acquire incomplete textures with limited representation and can not deal with large overall motions.

Some recent works [17]–[19] have shown that learning implicit neural representations of human body achieves remarkable reconstruction and reposing performance. However, one implicit model is only effective for one specific character, limiting its application prospects.

To summarize, the current schemes fail to tackle human posture imitation satisfactorily due to the following challenges: First, arbitrary posture manipulation of a human body belongs to non-rigid motions, which inevitably leads to enormous spatial layout and geometric changes. These variations impede high-quality image generation. Second, although it is simple to infer inherent textures of stationary clothes by existing methods, the character's motion inevitably produces folds and pleats on the clothes, which leads to deformations and distortions of clothes details reflected in images. Imitation limited to the 2D plane has difficulty estimating these deformations and distortions. It ultimately leads to undesired over-smoothing and poor details in target images, unrealistic distortions of joints and garments, or implausible texture alterations. Third, most methods are inferior at inferring, thus fail to synthesize textures of invisible parts rationally, which exceedingly limits their subsequent applications. Last, most of the existing approaches can only handle low-resolution situations. They fail to generate high-resolution and photo-realistic images without blur and artifacts.

In this paper, we propose Poxture (Human **Po**sture Imitation Using Neural Te**xture**), a novel human posture imitation method to address these challenges. Poxture adopts a parametric statistical human body model, SMPL [20], to estimate the character's shape and pose. Poxture extracts the source character's texture (appearance) from the source image with the help of an official UV mapping function. Then it converts the texture to a neural texture and fills missing parts of the texture during the conversion. The neural texture is warped onto the SMPL model with the source character's shape and the target character's pose. Then, we use a differentiable renderer with a weak-perspective camera to render the textured SMPL model. Note that the texture warped on the model is a neural texture, so the rendered image is a neural image with high-dimensional channels. Finally, a detail generator translates the neural image into an RGB image which is imposed to approach the target image. Thus, Poxture does not require any annotations for the source single-man video sequence and works with arbitrary appearance in a self-supervised manner. As depicted in Fig.2, Poxture can manipulate the character of the source image to imitate arbitrary poses performed by reference images without changing the shape of the character. Even if the source image does not supply any appearance information from the back side of the character (the last column), Poxture can generate reasonable and photo-realistic images with the ability to represent the back of the source character thanks to the learned neural texture estimator. We summarize the features of our method as follows:

A. Flexible

Thanks to the disentanglement of shape, posture, and appearance, our method can flexibly complete generation when there exist significant spatial changes and distortions between the source and the reference posture.

B. Accurate

Poxture employs 3D meshes to represent human emotions. Pose transfer between 3D meshes is more accurate and less ambiguous and thus can reduce misalignment compared to other intermediaries. Inspired by [21], we use neural texture to represent characters supervised by a pixel-level loss. Compared with traditional RGB storage as maps on top of the 3D mesh, the high-dimensional feature maps contain significantly more information.

C. Imaginative

We use an image translation network to infer the complete texture, which improves the scalability of the network. Moreover, we further encode human parsing images corresponding to the reference posture into the training pipeline, such that the detail completion network could achieve differentiated generation for different parts of the human body.

D. Universal

Poxture achieves a universal pipeline for human pose imitation with a single image. It can be easily applied to the images out of datasets since all appearance, shape, and pose information is directly acquired from the image regardless of the distribution of datasets.

We use the benchmark and evaluation protocol proposed in [12] to test our method. Experimental results show that our method can generate high-fidelity and photo-realistic results with high resolutions. It is proficient in migrating someone's appearance outside the dataset. In addition, based on the SMPL model and texture, our method can be easily extended to other tasks, such as garment transfer and novel view synthesis. It has influential scalability and robustness and can be competent for dramatically varied situations. The main contributions of our proposed method are summarized below:

- We introduce Poxture, a self-supervised and end-to-end method for human posture imitation. It builds a universal pipeline to implement high-resolution arbitrary motion transfer. Poxture fully disentangles the appearance, shape, and posture, easing the impact of dataset distribution. The neural texture Poxture acquired not only contains the RGB appearance but also encodes high frequency and physical attributes, which are the key to generating photo-realistic images. Besides, it provides the entire texture corresponding to the SMPL model of the source character, which is of great significance to related applications.

- Quantitative and qualitative experiments distinctly demonstrate the effectiveness of Poxture and show new state-of-the-art results on two challenging datasets, i.e., iPER [12] and DeepFashion [22].

II. Related Work

A. 3D Human Pose Estimation

3D skeleton estimation can be basically divided into two categories: one-step methods and two-step methods. One-step methods directly estimate 3D pose from the input image. While two-step methods output 2D skeleton locations in the first stage, and then upgrade these 2D keypoints to 3D locations by a learned dictionary of 3D skeleton [23]–[26] or regression [27]–[30].

In terms of the tasks based on a monocular RGB-only sequence, it is necessary to use extra anthropometric priors to improve the rationality of reconstruction. In this way, some researchers try to predict 3D pose through parametric human body model [20], [31], [32]. Bogo *et al.* [33] proposed the first method to automatically estimate the 3D pose of the human body as well as its 3D shape from a single unconstrained image. The model-based methods later got further developed and extended through the tight collaboration of regression- and optimization-based methods [32], [34]–[39]. Moreover, some other works [40], [41] estimated both 3D hand and body motions in a unified parametric model structure to make the reconstruction more vivid. Specifically, Chen *et al.* decompose the task into bone direction prediction and bone length prediction, from which the 3D joint locations can be completely derived. Since the bone lengths of a human skeleton remain consistent across time, this method significantly mitigates the depth ambiguity in many challenging poses. Temporal context can also be applied as supplementary information to ensure the accuracy and consistency of results in the time domains [42], [43].

B. Human Posture Imitation

Early works formulated human posture imitation as image-to-image translation to learn a mapping function from 2D skeletons or 2D dense poses to images [44]. For example, Shysheya *et al.* [45] used multi-view cameras to reconstruct the texture UV image of a person and learn a mapping from synthetic images to real images. Liu *et al.* [46] reconstructed and rendered a full 3D character avatar with a static pose from monocular video. However, the Achilles heel of translation-based methods is that they need to train specific models for different appearances.

Contemporarily, there are increasing studies which are based on the conditioned generative adversarial networks (CGAN) [5], [15], [47]–[52]. The intuition behind using CGAN is that GANs can take the concatenation of a source image and the source pose as the input and generate a realistic image using a reference pose. However, the pixel-wise aligned task is often troubled by the shape deformation between the source person image and target person image.

To mitigate this, recent works started to use 3D human poses to guide human image generation. Balakrishnan *et al.* [47] bridged the gap between the source and the target 2D key points with an affine transformation matrix and then generate the foreground and the background separately with a texture warping strategy. Moreover, Siarohin *et al.* [8] proposed PoseGAN, which computes an extensive affine transformation to solve the input-output misalignment caused by pose differences. On top of that, Li *et al.* [14] proposed to estimate 3D appearance flow from 2D key points and performed feature warping based on the learned transformations. Furthermore, Chen *et al.* [50] directly use the source poses and the target poses as conditions to guide image synthesis. However, the synthesized textures in these methods [3], [14], [50] suffer from being not consistent with the source in the real 3D space, resulting in incorrect modeling when position and scale change significantly. Zhang *et al.* [51] produce a target image by using the cross attention module on raw pixels and poses, in order to build the texture correlation between the source and the target domain. However, in this way, some valuable textures in the source image may not be transferred to the target synthesis image, which leads to an untrustworthy texture correlation. Ma et al. [52] use predicted segmented map to assist their synthesis. With segment maps and predicted poses, their model can build more accurate texture distributions according to different body parts. Since the target person's texture only appears to be the supervision of their model in training, Ma et al. [52] can only transfer motions to one particular person that the model has trained on. Zhang *et al.* [53] first synthesize a human parsing map corresponding to the target pose, and then utilize the poses and parsing maps from the source and the target to generate the final image. They also propose joint global and local per-region encoding and normalization to predict the reasonable style of clothing for invisible regions. These methods [51]–[53] use neural networks to build the correlation between human pose and appearance, but they perform poorly since they cannot fully utilize the source textures. In contrast, our method use the SMPL model as the 3D human body prior, samples all the textures from the source person and build a point-by-point map from the source texture to the 3D human model, which build a strong correlation between human body and texture. Differ from human appearance transfer [15], [54], which aims at transferring a source character's texture on a target image, resulting in an image with the source character's appearance and a target character's shape and posture, we focus on generating photo-realistic images with the source character's appearance, the source character's shape and the target character's pose.

C. Neural Texture

Thies *et al.* [55] first proposed Neural Textures in contrast to the traditional texture which we term it the appearance texture. This is a rendering approach that uses meshes instead of 3D points to embed learnable feature vectors to optimize a neural texture in conjunction with a deferred neural renderer and generate high quality facial reenactments. Then Thies et al [56] proposed a novel light-weight neural rendering network based on neural textures to re-render human faces. Aliev et al further [57] presented a point-based approach for modeling

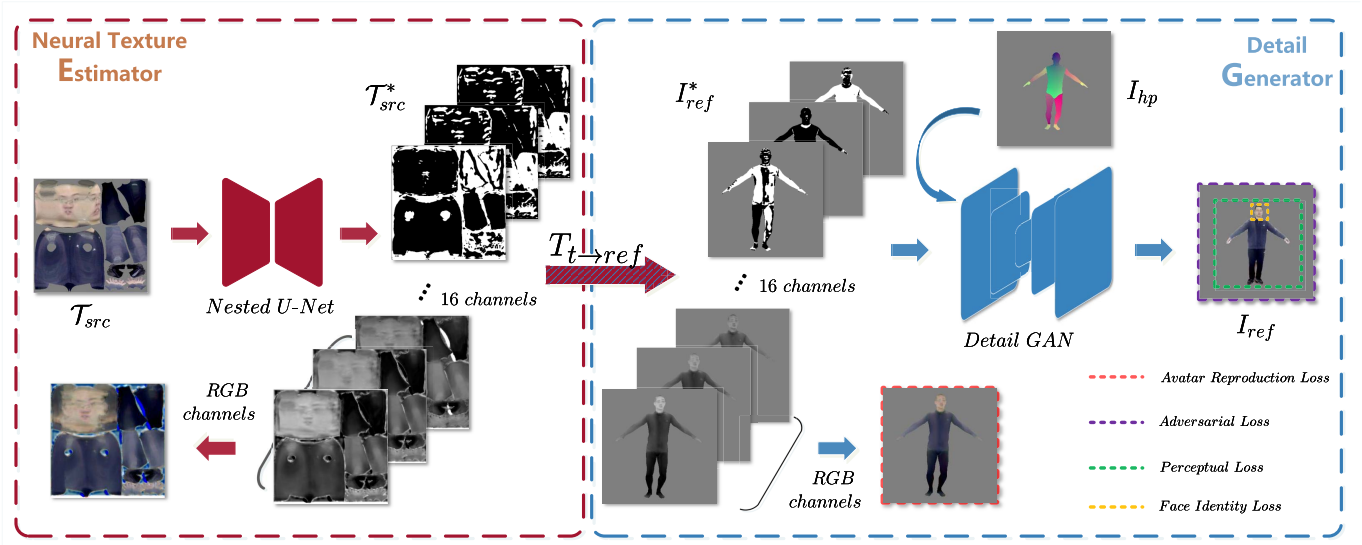

Fig. 1. The overview of our training pipeline. The colors of the dash lines at the bottom right corner indicate the correspondence between the loss functions and the regions where they are computed on. Given a source texture $\mathcal{T}_{src}$ extracted from the source image, our neural texture estimator E fixes irrational parts and converts it to a full 16 channels texture feature map $\mathcal{T}^*_{src}$. The first three channels of $\mathcal{T}^*_{src}$ are explicitly supervised by RGB semantic information. The rest are implicit texture feature channels (we normalize their intensities for better visualization). Given $\mathcal{T}^*_{src}$ and a reference pose, we can re-pose the source SMPL model to imitate the target pose and calculate the transform matrix $T_{t\to ref}$. Then $T_{t\to ref}$ is used to sample $\mathcal{T}^*_{src}$ by bi-linear interpolation and get a neural image I^*_{ref}. Finally, the detail GAN receives I^*_{ref} and a face map I_{hp} from the source SMPL model with reference pose as input and converts them to a realistic image.

the appearances of real scenes, which also learns neural descriptors of surface elements jointly with the rendering network but uses point-based geometry representation and thus avoids the need for surface estimation and meshing. Moreover, Sarkar *et al.* [54] tried to take advantage of the learned feature map including fine-scaled geometry and clothing textures on the task of texture-based human re-rendering and achieved sharper results compared to other natural choices. However, they utilized DensePose to predict the correspondence between pixels and 3D meshes, resulting the consequence of misalignment. Our Poxture remedies the problem with pseudo 3D human priors.

III. Method

In this section, we describe our Poxture model. The overall pipeline of our model is shown in Fig. 1. Poxture can be divided into three parts: ***Preliminary***, ***Neural Texture Estimator E*** and ***Detail Generator G***. Given a source image I_{src} with source appearance and a reference image $\hat{I}$ with target posture in size $H \times W$, the Preliminary part predicts camera parameters and the 3D mesh of I_{src} and $\hat{I}$. Then it generates pseudo texture with the help of the official UV mapping function. Neural Texture Estimator estimates the complete texture of I_{src} represented by the neural texture. Detail Generator samples and warps the estimated neural texture $\mathcal{T}_{src}$ on the source SMPL model and renders the textured model into a pseudo image. Then it translates this pseudo image to an actual RGB image with attached details. We describe details of these modules in the following subsections.

A. Preliminary

In this paper, we use SMPL to reconstruct the 3D human body. SMPL can drive the human body model through shape and posture parameters and ultimately disentangles shape and posture. It outputs a 3D mesh consisting of vertices and faces. The SMPL mesh can be built by a differentiable function:

$$Mesh = M(\theta, \beta) \in \mathbb{R}^{3K+10} \tag{1}$$

where K is the number of joints. $Mesh$ is a triangulated mesh with $N_{vex} = 6,890$ vertices and $N_{face} = 13,776$ faces. Here, one face consists of three vertices in a clockwise or anticlockwise direction. The pose parameters $\theta \in \mathbb{R}^{K\times 3}$ symbolize the rotations of $K-1$ joints concerning their parent joints and global body rotation of the root joint in the form of the axis-angle. The shape parameters $\beta \in \mathbb{R}^{10}$ are the first 10 orthogonal bases of PCA feature space.

SMPL applies corresponding parameters to control the shape and pose, respectively. This property allows us to control the SMPL model without changing the character's shape by adjusting the pose parameters. In this article, we transfer the reference SMPL pose parameters to the source SMPL mesh M_{src} and get the desired mesh $\hat{M}_{src}$ with both source shape and reference posture.

To get the SMPL representation of characters in I_{src}, we use framework SPIN proposed in [58] as the 3D pose estimator. SPIN provides M_{src} and camera parameters $c = (t, s) \in \mathbb{R}^3$. Here, $t \in \mathbb{R}^2$ are the translation parameters and $s \in \mathbb{R}$ are the scale parameters of the weak perspective camera model utilized by Poxture.

Fig. 3 shows the general procedure of texture extraction. To extract the texture of source characters, we use the camera with parameters c_{src} to project vertices v_{src} of M_{src} to 2D image plane through a rasterization based method. In this way, we get the 2D coordinates of the predicted SMPL model's vertices on the image plane. Then, each face of the SMPL model onto the image plane is determined by the

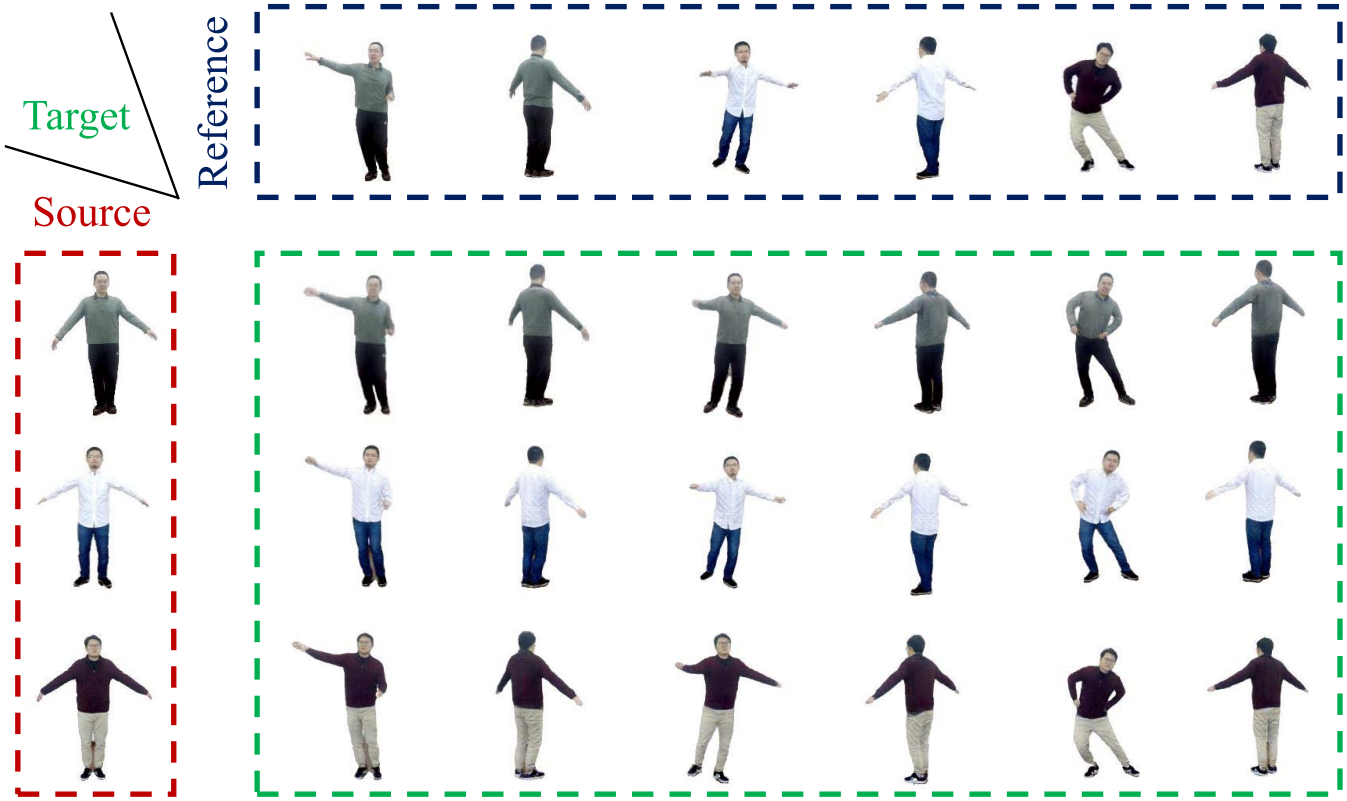

Fig. 2. The qualitative results of our method - Poxture. Given a source character and a target pose, our method (Poxture) can generate the source character with the target pose in a weakly supervised way. Poxture employs a 3D parametric human model to use the 3D prior hidden in the source image entirely. In this way, we can control the source character to move arbitrarily and reasonably. Besides, Poxture enables strong inference ability. Even if the texture provided by the source image is insufficient to cover the target pose (the last column), it can correctly infer the desired texture and reasonably warp it to the corresponding position.

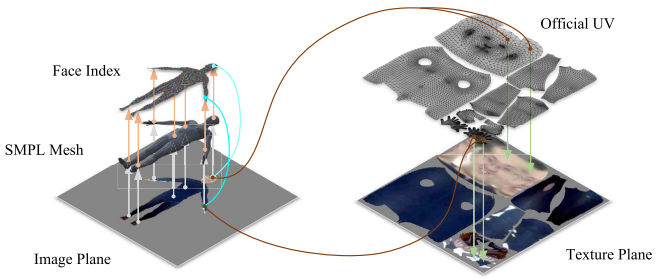

Fig. 3. The procedure of extracting texture from a single image.

three vertices' projected coordinates. Besides, with the help of c_{src}, the barycentric weight of coordinates on both the image plane and the texture plane in their corresponding face can be calculated.

By finding related coordinates between the source image and the texture plane, we can obtain a face index map I_{hp} representing human parsing, whose value for each pixel represents the face index in M_{src}. We also calculate the sample matrix $T_{src\to t} \in \mathbb{R}^{H\times W\times 2}$ of $p_{i,j}$ by its barycentric weight. This process determines a mapping from faces on I_{src} to faces on the texture plane. Finally, we use $T_{src\to t}$ to sample I_{src} and acquire the texture image $\mathcal{T}_{src}$. Note that this mapping function is reversible and differentiable; thus, we can use the reversed transformation matrix $T_{t\to ref}$ to warp textures to reference posture.

$$\mathcal{T}_{src} = T_{src\to t}(I_{src}) \tag{2}$$

Unfortunately, $\mathcal{T}_{src}$ produced in this way represents a "double-faced" person. We term such "double-faced" texture as pseudo texture. Since 2D RGB images only provide information on one side, based on sampling, information on the invisible side would be consistent with information in the visible region, resulting in duplicated textures on both sides. For example, suppose that the source image only provides the frontal view of a human. When we sample the head, we sample a face and record it on the texture plane. When we sample the posterior cranium, we sample the face again; there will be two same faces on the resulting texture. There are two possible solutions to address this issue. One is to calculate the actual image orientation according to the SMPL model's normals and record invisible parts as background. We call the texture extracted in this way partial texture. However, the position of visible parts inputted to the model varies during training, which results in unstable training procedures. The imitation results suffer from artifacts inevitably. The other is to use the pseudo texture as input directly. Then a specialized network trained with a self-supervised strategy is applied to correct its

irrational parts. This network can understand the varieties and relations between different regions in texture space. At the same time, this network can preserve detailed information from the source image. We adopt the second solution, and the detailed comparison results between the two methods above are introduced in Sec. V-E1.

B. Neural Texture Estimator

In this section, we introduce the Neural Texture Estimator E. Poxture converts $\mathcal{T}_{src}$ to the neural texture $\mathcal{T}^*_{src}$ using E which efficiently encodes the entire appearance of the source character in normalized texture space. We assign two tasks for the estimator E: First, E should be able to transfer the pseudo texture mentioned in Sec. III-A to a high dimensional space where all source characters' appearance can be fully expressed. Second, E is supposed to reproduce a reasonable texture from the inputting pseudo texture. Since the pseudo texture derives from a "double-faced" character, some texture contains wrong information. Similar to [59], we use a nested U-net introduced by [60] as the estimator E to correct the pseudo texture. The nested U-net is a robust image translation architecture. The original structure of nested U-net is pretty bulky, so we prune it at the second level to form our E module since it behaves almost the same as the more complicated L_3 and L_4 (L_i denotes a U-net pruned at level i).

$$\mathcal{T}^*_{src} = E(\mathcal{T}_{src}) \tag{3}$$

E can fix the pseudo texture $\mathcal{T}_{src}$ and finally prepare a neural texture corresponding to I_{src}. Inspired by [21], $\mathcal{T}^*_{src}$ generated by E owns 16 channels, of which the first 3 channels enable RGB semantic information. Subsequent ablation experiments show that our implementation enables shorter training procedures and achieves better results with the image generated more photo-realistic than the method without RGB channels.

C. Detail Generator

Since the complete appearance of the source character is stored in $\mathcal{T}^*_{src}$, the human posture imitation task can be completed naturally by "dressing" the neural texture on the source SMPL model with the reference posture.

First, a reverse approach introduced in Sec. III-A is employed to generate the transformation flow $T_{t\to ref} \in \mathbb{R}^{H\times W\times 2}$ from $\mathcal{T}^*_{src}$ to a textured SMPL model. Then we use a differentiable renderer introduced by [62] to render the 3D model into a pseudo image $I^*_{ref} \in \mathbb{R}^{H\times W\times 16}$.

$$I^*_{ref} = f(T_{t\to ref}(\mathcal{T}^*_{src}, M_{src})) \tag{4}$$

Here, M_{src} is the estimated source SMPL model, and f denotes the differentiable renderer.

I^*_{ref} is sampled from $\mathcal{T}^*_{src}$, hence I^*_{ref}'s first 3 channels can be interpreted as RGB. SMPL models do not capture clothes, so the first 3 feature channels of I^*_{ref} looks like drawing colors on a human body surface, which is different from the realistic image. To address this issue and convert I^*_{ref} from a neural image to a RGB image, we design a Detail Generator G. Current researches have shown promising results on reconstructing 3D clothed human models in a general way [63]–[65]. However, these approaches rely on complex networks, but yielding coarse results. In contrast, our G is a small but effective network. To be more concrete, it is a conditional GAN similar to Pix2Pix [66]. G contains a 4-level U-net based generator with 19 (16+3, 16 channels are designed for neural texture and three channels are for a face index map) input channels and a fully convolutional patch discriminator D to discriminate between I_{ref} and $\hat{I}$. G gathers features from I^*_{ref} and translate I^*_{ref} into the RGB domain. In addition, we concatenate the face map I_{hp} of $\hat{M}_{src}$ and I^*_{ref} as the input of G. In this way, human parsing information is encoded and can help the detail GAN accurately recognize the different parts of the human body.

$$I_{ref} = G(I^*_{ref}, I_{hp}) \tag{5}$$

IV. Overall Loss Function for Poxture

Since Poxture consists of several modules, its loss function is formed by four terms: Avatar Reproduction Loss, Adversarial Loss, Perceptual Loss [67] and Face Identity Loss.

A. Avatar Reproduction Loss

Inspired by [21], we employ an avatar reproduction loss $\mathcal{L}_a$ to enforce that the first 3 feature channels of I^*_{ref} represent the color of $\hat{I}$. This loss function is designed specifically for E, and its formulation is given as follows:

$$\mathcal{L}_a = \|I^*_{ref}[0:3] - \hat{I}\|_1 \tag{6}$$

Here, $I^*_{ref}[0:3]$ is the first 3 channels of I^*_{ref}.

B. Adversarial Loss

We use an adversarial loss $\mathcal{L}_{adv}$ to push the distribution of generated images to the distribution of real images. Here we use the $LSGAN$ [68] loss to act as $\mathcal{L}_{adv}$ with I_{ref} and its corresponding ground-truth $\hat{I}$. $\mathcal{L}_{adv}$ works on G, and it is given by:

$$\mathcal{L}_{adv} = \sum D(I_{ref}, \hat{I})^2 \tag{7}$$

C. Perceptual Loss

We adopt a perceptual loss $\mathcal{L}_{vgg}$ to enforce the generated images and the reference images to be close in VGG [69] feature space. It works on the whole pipeline. The formulation is shown as follows:

$$\mathcal{L}_{vgg} = \sum_{i=1}^{n} \alpha_i \|V_i(I_{ref}) - V_i(\hat{I}_{ref})\|_1 \tag{8}$$

Here V represents a pre-trained VGG-19 [69] on ImageNet [70]. $V_i(\cdot)$ denotes the i^{th} channel feature and α_i is the weight. In all our experiments, we use i in $2, 7, 12, 21, 30$ and corresponding weights are $1/32, 1/16, 1/8, 1/4, 1$.

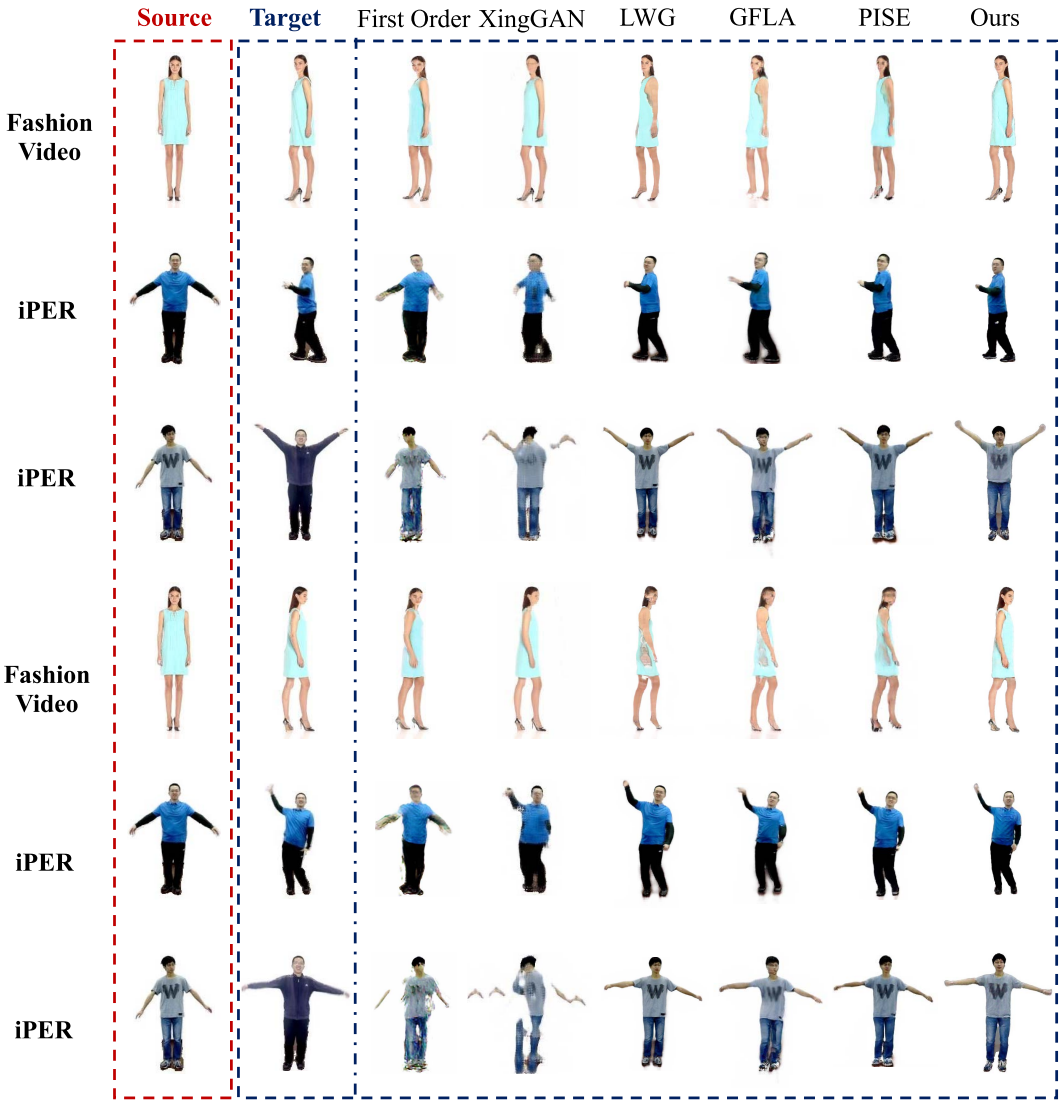

Fig. 4. A qualitative comparison with states-of-the-arts. From left to right are the results of First Order [4], XingGAN [3], LWG [12], GFLA [61], PISE [53] and ours, respectively. The first and fourth columns show the source images from different datasets to driving other frames. The second and seventh columns illustrate the reference frames with poses we want to generate. In order to better compare the human posture imitation ability with others, in the first, second, fourth and fifth rows, we show the results under the self-imitation setting, i.e., the source and the reference frames in a row come from the same video. In the third and sixth row, we show the results under a cross-imitation setting, i.e., the source and the reference frames come from different persons. It can be found that our results are more delicate, natural, and reasonable than others.

D. Face Identity Loss

The face has more high-frequency information than other body parts, so it is very tough for the network to generate a reasonable and clear face. To ensure E and G can preserve the face identity, we employ a face identity loss $\mathcal{L}_{face}$ to assist the generation of face regions. $\mathcal{L}_{face}$ works on the whole pipeline. It is computed by:

$$\mathcal{L}_{face} = \sum_{i=1}^{n} \lambda_i \| S_i(I_f) - S_i(\hat{I}_f) \|_1 \tag{9}$$

Here, S is a pre-trained SphereFaceNet-20 [71]. $S_i(\cdot)$ denotes the i^{th} channel feature of the SphereFaceNet-20 and λ_i is the weight. I_f and $\hat{I}_f$ denote the face regions of I_{ref} and $\hat{I}$ respectively.

Hence, the overall loss is:

$$\mathcal{L}_{Poxture} = \mathcal{L}_{adv} + \lambda_{vgg} \cdot \mathcal{L}_{vgg} + \lambda_{face} \cdot \mathcal{L}_{face} + \lambda_a \cdot \mathcal{L}_a \tag{10}$$

We train all our experiments with $\lambda_{vgg} = 0.5$, $\lambda_{face} = 0.1$ and $\lambda_a = 0.1$.

V. Experiments

In this section, we first introduce the experimental setup in detail. Then we compare our results with some state-of-the-art approaches and present qualitative results as is shown in Fig. 4. Finally, we conduct ablation experiments to show the effect of each component or design choice in our proposed architecture.

TABLE I

ONE-SHOT HUMAN POSTURE IMITATION RESULTS OF DIFFERENT METHODS ON THE IPER AND FASHION VIDEO DATASETS. THE BEST RESULT AND SECOND BEST RESULT UNDER EACH METRIC ARE SHOWN IN BOLD AND UNDERLINE RESPECTIVELY. "OURS" AND "OURS*" INDICATE OUR IMITATION MODEL IS TRAINED ON IPER AND THE MIXTURE OF IPER AND FASHION VIDEO, RESPECTIVELY

	iPER				Fashion Video			
	SSIM ↑	PSNR ↑	FID↓	LPIPS↓	SSIM↑	PSNR↑	FID↓	LPIPS↓
First Order [4]	0.867	17.522	110.254	0.155	0.901	27.193	32.785	0.058
XingGAN [3]	0.813	17.534	70.486	0.148	0.891	26.074	35.354	0.076
LWG [12]	0.866	26.302	38.669	0.068	0.912	28.315	<u>29.625</u>	0.045
GFLA [62]	0.859	24.271	35.075	0.072	0.894	27.513	30.213	0.071
PISE [54]	0.883	27.312	**30.432**	0.070	0.907	30.381	30.125	0.051
Ours	**0.897**	<u>28.030</u>	32.509	<u>0.065</u>	<u>0.918</u>	<u>32.868</u>	30.119	<u>0.042</u>
Ours*	<u>0.892</u>	**28.313**	<u>32.011</u>	**0.064**	**0.921**	**32.872**	**29.421**	**0.041**

A. Dataset

We mainly conduct experiments on Fashion Video [22] and iPER [12]. Fashion Video has 500 training and 100 test videos, each containing around 350 frames. Each video shows a different single female subject wearing a different cloth with white background and presents the action starting with her front and slowly turning in a circle. iPER has 30 subjects of different shapes, heights, and genders, each wearing different clothes. Some subjects might have multiple clothes, and there are 103 clothes in total. An A-pose video and a video with random actions are provided for each subject. The whole dataset contains 206 video sequences with 241,564 frames, divided into training and test set at the ratio of 8:2.

B. Quantitative Evaluation

For evaluation metrics, we use Peak Signal-to-Noise Ratio (PSNR), Inception Score (IS), Structural Similarity Index Measure (SSIM), and Learned Perceptual Similarity (LPIPS). SSIM and PSNR are similarity metrics, the higher the better. FID and LPIPS are distance metrics, the lower the better. We compare our methods with some state-of-the-art posture imitation methods, including First Order Motion Model (First Order) [4], XingGAN [3], Liquid Warping GAN with Attention (LWG) [12], Global-Flow-Local-Attention (GFLA) [61] and PISE [53]. Notice that, since our GPU memory is limited, we train XingGAN [3] and PISE [53] with 256×256 images and train other methods with the exact resolution as ours at 512×512. All the images are cropped by the method mentioned in [72] and are padded with white pixels to be the original size. Since Poxture can perform human pose imitation with only SMPL pose parameters, we use two images from one character with different poses for quantitative evaluation, termed self-imitation. In addition, we train our Poxture on a mix of iPER and FashionVidieo datasets, termed ours*. The quantitative comparison result is shown in Table I. It can be easily found that in most cases, our method achieves better human posture imitation results. Our results achieve the best performance in terms of SSIM, PSNR, and LPIPS, which indicates that our model not only generates more realistic images but also keeps the harmony of appearance and shape. Our model fails to outperform state-of-the-art in terms of FID. This is probably because our model pays much attention to transferring detailed textures from the source domain to the target domain. But these transferred textures may look inappropriate on the target pose. This is an limitation need to be address in future work. As the quantitative results shown in Table I, training our model on the mixed dataset leads to slightly better performance. This phenomenon is also observed from the results of LWG.

C. Qualitative Evaluation

To intuitively evaluate the quality of generated images, we designed a qualitative evaluation strategy. We choose a frontal frame from a video in iPER or Fashion Video as the source image and choose another video from the same dataset as the driven video. We apply the methods mentioned above to generate the target video and fetch frames to compare the source image and the driven video as inputs. Since we acquire the SMPL model and texture of the character in the given image, it is flexible for us to conduct imitation. We can either transfer the texture of the source character to the target character's SMPL model or manipulate the source SMPL model to imitate the target SMPL model and warp the source texture back. The latter strategy is more in line with the definition of human pose imitation, and we term it as cross-imitation in contrast with the self-imitation. The qualitative comparison is shown in Fig.4.

As the figure shows, our results are more delicate and natural compared to others. Compared with First Order Motion Model (based on local affine transformation) [4] and XingGAN (based on image transfer) [3], our method acquires a more prior solid knowledge of humans due to the reconstruction of 3D meshes. Thus it is better at the specific human imitation task. For some appearance information beyond the source images, our results show a better complementary ability to generate realistic images, such as hands in row 1, 4 and the side view in row 2, 5, particularly compared with LWG with the "double-faced person" problem. GFLA [61] can transfer the detailed texture from the source image to target image. However, it has difficulty in generating reasonable results for the invisible regions of the source image. With human parsing maps, PISE [53] can accurately represent the target pose and shape. However, when facing large movements,

TABLE II

QUALITATIVE RESULTS ON DIFFERENT OCCLUSION LEVELS. AS THE OCCLUSION LEVELS BECOMES HIGHER, OUR METHOD SUFFER FROM PERFORMANCE DEGRADATION, BUT THE DEGRADATION IS MARGINAL AND ACCEPTABLE

Levels	SSIM ↑	PSNR ↑	FID↓	LPIPS↓
Standard	0.897	28.030	32.509	0.065
Distorted	0.861	27.661	36.918	0.076
Occlusive	0.855	25.031	48.937	0.091

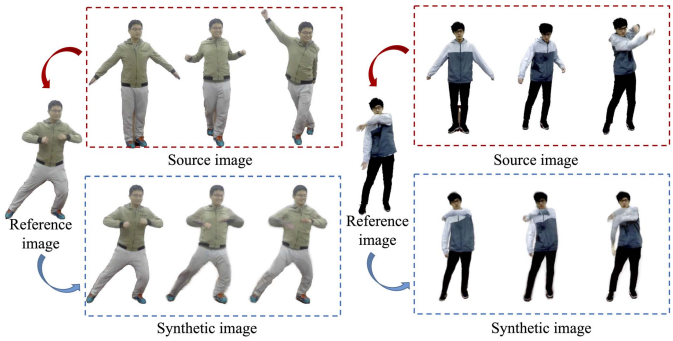

Fig. 5. Qualitative results of the occlusion experiment. In each group of source images, we show a standard pose, a distorted pose and an occlusive pose, respectively. As observed from the synthetic images, even with severe self-occlusion in the source images, our Poxture can still synthesize high fidelity images.

PISE [53] cannot keep the details of the source character well. In comparison, our results maintain more texture details inherited from the source image, even with large movements. Moreover, our results combine the target pose and the source appearance more naturally, making the imitation results more reasonable. With the help of texture implementation, our imaginative network can handle overlapping and invisible data. Compared to the Fashion Video dataset, test videos in iPER have more various and complicated actions. This specific situation requires networks to have a better generalization ability, which First Order Motion Model [4] and XingGAN [3] are not holding. In row 3 and 6, we demonstrate that our method can manipulate the source character without shape and appearance collapse. Note that we retain the detailed texture information, for instance, the "W" of the T-shirt in row 3 and 6.

D. Occlusion Analysis

The self-occlusion in human bodies is a common challenge in human pose imitation. To verify the robustness of our method against self-occlusion, we conduct experiments under different levels of occlusion. This can be controlled by tuning the human pose in the source image: 1) Standard pose, which provides a complete frontal view; 2) Distorted pose, which has slight self-occlusion due to small limb movements; 3) Occlusive pose, which suffers from severe self-occlusion due to large limb movements. One can observe that the occlusion levels of the three types of human poses become higher and higher.

Table II and Fig. 5 report quantitative and qualitative results of the occlusion experiment. Our Poxture can fix distorted textures and generate unobserved areas reasonably with only a minor performance downgrade. As mentioned above, Poxture builds a bridge between pose and appearance through 3D human geometry. The explicit and complete human structure representation, along with human parsing encoding, assists the network in better understanding the semantic information of the human body and enhances its generation and completion ability. In addition, the pseudo texture provides symmetrical information compared to the partial texture, thus enhancing the performance with high occlusion levels in source images.

E. Ablation Study

In this subsection, we explore the impact of each component or design choice in our pipeline, including the comparison between using the pseudo texture and using the partial texture as the input for the Neural Texture Estimator mentioned in Sec. III-A, whether to learn explicit RGB channels (the first three) in the neural texture, the influence of different channel numbers of the neural texture, whether to use human parsing encoding to assistant the detail GAN for target image generation. All these ablation studies are conducted on the iPER dataset with the same training strategy and evaluation method.

1) Pseudo Texture (PT): To verify the importance of using the pseudo texture as the input for the Neural Texture Estimator, we conduct a study that using the partial texture as the input instead.

2) Explicit RGB Channels (EC): We conduct an ablation study on RGB channels of the neural texture to verify their effectiveness. To this end, we train a Poxture model without RGB layers in $\mathcal{T}^*_{src}$. The semantic supervision attached on the RGB channels, i.e., the first three layers of $\mathcal{T}^*_{src}$, is assigned by $\mathcal{L}_a$. So we discard $\mathcal{L}_a$ during the training period with other modules unchanged.

3) Human Parsing Encoding (HPE): We conduct an ablation study to show the effectiveness of human parsing encoding for high-quality target image generation. To do so, we exclude I_{hp} from the input for the detail GAN G, which reduces the number of its input channels to 16.

4) Feature Channel Numbers: Most current texture extraction methods [12], [13], [73]–[75] directly warped the RGB channels from a source image onto a desired human body. To demonstrate the effectiveness of our implicit texture representation and investigate the influence of different channel numbers of the neural texture, we design an ablation study to verify different designs for the channels of the neural texture $\mathcal{T}^*_{src}$: 4.1): $\mathcal{T}^*_{src}$ has only the RGB channels, termed as **F3**; 4.2): $\mathcal{T}^*_{src}$ has 5 neural texture channels and the RGB channels, termed as **F8**; 4.3): $\mathcal{T}^*_{src}$ has 13 neural texture channels and the RGB channels, which we adopt in this paper, termed as **Full**; 4.4): $\mathcal{T}^*_{src}$ has 21 neural texture channels and the RGB channels, termed as **F24**.

As the ablation results shown in Table III, the neural textures with more feature channels is richer intermediate representation, and thus produces more realistic images than simply using appearance texture. This is also evidenced by qualitative results shown in Fig. 7, i.e., the images generated

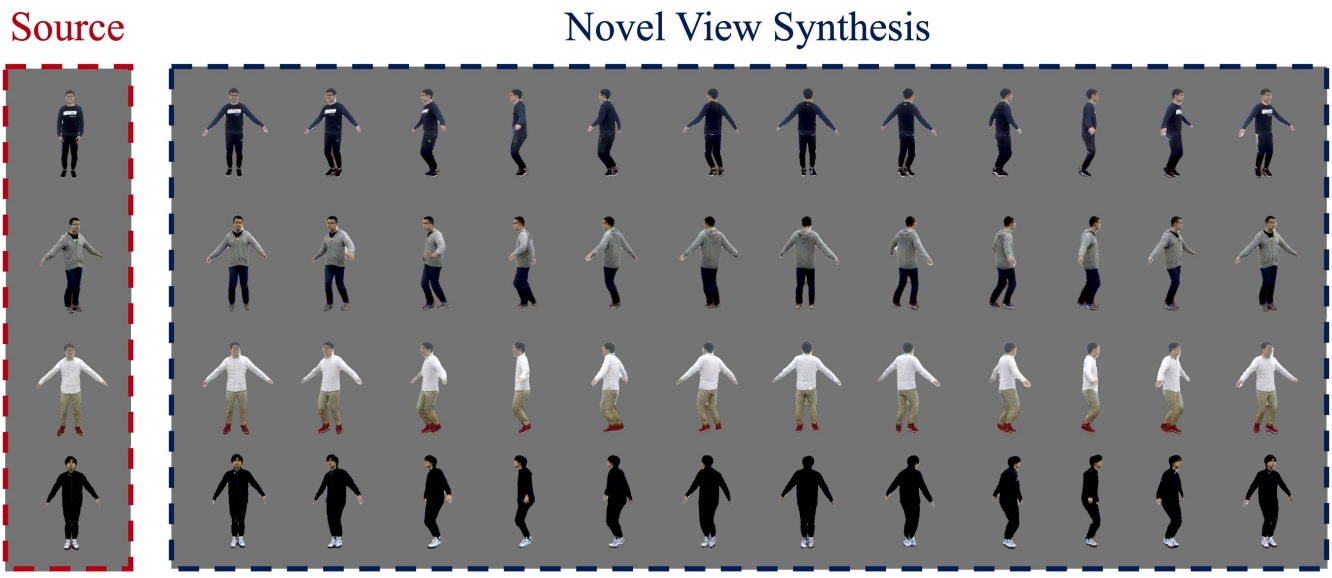

Fig. 6. Examples of our Poxture novel view synthesis performances on the iPER dataset with 512 × 512 resolution. Our method changes the sampling region of the latent space to get neural images from different views to ensure the posture is the same. We can observe that our method can infer the invisible parts of human bodies, even if the front and the back are quite different.

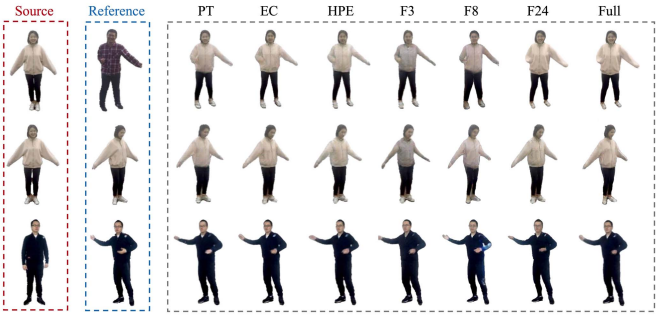

Fig. 7. Qualitative results of different variants of our method. From left to right: source image, reference image, images generated from: Pseudo Texture (PT), Explicit RGB Channels (EC), Human Parsing Encoding (HPE), only RGB channel (F3), five feature channels with RGB channels (F8), twenty-one feature channels with RGB channels (F24), thirteen feature channels with RGB channels (Full).

TABLE III

QUALITATIVE RESULTS OF THE ABLATION STUDY ON THE TEST SET OF IPER WITH 512 × 512 RESOLUTION

Method	SSIM ↑	PSNR ↑	FID↓	LPIPS↓
PT	0.867	23.612	56.144	0.135
EC	0.881	27.641	35.411	0.073
HPE	0.813	22.113	50.198	0.169
F3	0.789	20.167	110.081	0.231
F8	0.851	25.508	42.096	0.108
F24	**0.901**	28.012	33.106	**0.065**
Full	0.897	**28.030**	**32.509**	**0.065**

by our neural textures are photo-realistic, with more details and more reasonable color distributions. However, as the channel numbers of the neural texture increase, our imitation model becomes heavier, which results in higher computational cost during inference. However, its imitation performance is not improved accordingly. By weighing time efficiency and imitation performance, we set the number of neural texture channels to be 13 (with 3 RGB channels) as our full imitation model.

F. Extensions

Our method samples a complete latent space to encode the appearance of the corresponding human body so that we can extend our work to novel view synthesis tasks without further training. Given a desired view direction, we first calculate the rotation R and the drift d. Then we calculate the novel view 3D mesh M_n by multiplying R and adding d, $M_n = M_{src}R + d$ to the original 3D mesh of the source human. In this way, we reconstruct a 3D mesh in the desired direction. Finally, the neural texture obtained from E is warped on this model and then fed into the detail generator to generate the target image from the novel view. As shown in Fig. 6, our method can achieve impressive performance on this novel view synthesis task.

VI. CONCLUSION AND FUTURE WORK

In this paper, we present Poxture, a novel human posture imitation method that enables accurate human motion capturing and human texture mapping. Thanks to the complete disentanglement of appearance, shape, and posture, Poxture realizes the genuine human posture imitation, generating images with source appearance, source body shape, and target body pose with a single image or SMPL parameters provided. Our model achieves practical improvements on several challenging datasets compared to existing state-of-the-art methods and produces decent results. This method will enable the application of future creative content creation.

While Poxture is free from human shape, we acknowledge that this method has some limitations: 1) Poxture heavily relies on the results of the SMPL regressor. The regression accuracy will directly affect the downstream process. If the SMPL model of the source character or the target character has a significant estimation error, it will lead to a strange human texture or weird posture. In the future, we can jointly learn the SMPL regressor and our E module and G module within our pipeline to guarantee the precision of the SMPL model. 2) Our approach is bottle-necked by the parametric human body model, limiting the expression and generalization capacity on off-body dressing such as sun hat and long skirt. A deformable parametric human body model or neural radiance field is an exciting direction for future work.

REFERENCES

[1] F. Xu *et al.*, "Video-based characters: Creating new human performances from a multi-view video database," in *Proc. SIGGRAPH*, Aug. 2011, pp. 1–10.

[2] K. Li *et al.*, "SPA: Sparse photorealistic animation using a single RGB-D camera," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 771–783, Apr. 2016.

[3] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 717–734.

[4] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2019, pp. 1–11.

[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[6] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8620–8628.

[7] B. AlBahar and J.-B. Huang, "Guided image-to-image translation with bi-directional feature transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2019, pp. 9016–9025.

[8] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3408–3416.

[9] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[10] H. Zheng, L. Chen, C. Xu, and J. Luo, "Pose flow learning from person images for pose guided synthesis," *IEEE Trans. Image Process.*, vol. 30, pp. 1898–1909, 2021.

[11] L. Yang *et al.*, "Towards fine-grained human pose transfer with detail replenishing network," *IEEE Trans. Image Process.*, vol. 30, pp. 2422–2435, 2021.

[12] W. Liu, Z. Piao, Z. Tu, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN with attention: A unified framework for human image synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 7, 2021, doi: 10.1109/TPAMI.2021.3078270.

[13] J. S. Yoon, L. Liu, V. Golyanik, K. Sarkar, H. S. Park, and C. Theobalt, "Pose-guided human animation from a single image in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15039–15048.

[14] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3693–3702.

[15] N. Neverova, R. A. Güler, and I. Kokkinos, "Dense pose transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 11207, 2018, pp. 128–143.

[16] D. Wei, X. Xu, H. Shen, and K. Huang, "C2F-FWN: Coarse-to-fine flow warping network for spatial-temporal consistent motion transfer," 2020, *arXiv:2012.08976*.

[17] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Occupancy flow: 4D reconstruction by learning particle dynamics," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5378–5388.

[18] Z. Yang *et al.*, "S3: Neural shape, skeleton, and skinning fields for 3D human modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13284–13293.

[19] S. Peng *et al.*, "Animatable neural radiance fields for modeling dynamic human bodies," 2021, *arXiv:2105.02872*.

[20] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Oct. 2015.

[21] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[22] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "Dwnet: Dense warp-based network for pose-guided human video generation," in *Proc. BMVC*, 2019, p. 51.

[23] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1446–1455.

[24] H.-Y.-F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4364–4372.

[25] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4966–4975.

[26] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3D human pose estimation with bone-based pose decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 198–209, Feb. 2021.

[27] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9915, 2016, pp. 156–169.

[28] H. Fang, Y. Xu, W. Wang, X. Liu, and S. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *Proc. AAAI*, 2018, pp. 6821–6828.

[29] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2659–2668.

[30] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2621–2630.

[31] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape completion and animation of people," in *Proc. ACM SIGGRAPH*, 2005, pp. 408–416.

[32] G. Pavlakos *et al.*, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10975–10985.

[33] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 9909, 2016, pp. 561–578.

[34] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6050–6059.

[35] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *Proc. 3DV*, Sep. 2018, pp. 484–494.

[36] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 459–468.

[37] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[38] G. Varol *et al.*, "BodyNet: Volumetric inference of 3D human body shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 20–36.

[39] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2148–2157.

[40] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 20–40.

[41] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration," 2020, *arXiv:2008.08324*.

[42] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3D human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3395–3404.
[43] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5252–5262.
[44] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5932–5941.
[45] A. Shysheya *et al.*, "Textured neural avatars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2387–2397.
[46] L. Liu *et al.*, "Neural rendering and reenactment of human actor videos," *ACM Trans. Graph.*, vol. 38, no. 5, p. 139, Oct. 2019.
[47] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8340–8348.
[48] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.
[49] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses for pose-based human image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 118–126.
[50] B. Chen, Y. Zhang, H. Tan, B. Yin, and X. Liu, "PMAN: Progressive multi-attention network for human pose transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 302–314, Feb. 2021.
[51] P. Zhang, L. Yang, X. Xie, and J. Lai, "Lightweight texture correlation network for pose guided person image generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4584–4598, Nov. 2021.
[52] F. Ma, G. Xia, and Q. Liu, "Spatial consistency constrained GAN for human motion transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 730–742, Mar. 2021.
[53] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "PISE: Person image synthesis and editing with decoupled GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7982–7990.
[54] K. Sarkar, D. Mehta, W. Xu, V. Golyanik, and C. Theobalt, "Neural re-rendering of humans from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 12356, 2020, pp. 596–613.
[55] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, p. 66, Jul. 2019.
[56] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 716–731.
[57] K.-A. Aliev, A. Sevastopolsky, M. Kolos, D. Ulyanov, and V. Lempitsky, "Neural point-based graphics," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 696–712.
[58] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2252–2261.
[59] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 84–93.
[60] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, 2019.
[61] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7690–7699.
[62] S. Liu, T. Li, W. Chen, and H. Li, "A general differentiable mesh renderer for image-based 3D reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 50–62, Jul. 2020.
[63] E. Corona, A. Pumarola, G. Alenyà, G. Pons-Moll, and F. Moreno-Noguer, "SMPLicit: Topology-aware generative model for clothed people," 2021, *arXiv:2103.06871*.
[64] Q. Ma *et al.*, "Learning to dress 3D people in generative clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6469–6478.
[65] B. Lal Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3D human reconstruction," 2020, *arXiv:2007.11432*.
[66] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[67] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.
[68] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2947–2960, Dec. 2019.
[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[70] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[71] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.
[72] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[73] G. Pavlakos, N. Kolotouros, and K. Daniilidis, "TexturePose: Supervising human mesh estimation with texture consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 803–812.
[74] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12135–12144.
[75] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *Proc. 3DV*, 2019, pp. 643–653.

Chen Yang received the B.S. degree from the Department of Instrument Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. He is currently pursuing the Ph.D. degree with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University. His research interests include neural rendering and neural reconstruction.

Shun-Yu Yao received the B.S. degree from the Department of Mechanical Engineering, Xi'an Jiaotong University, and the M.S. degree from the Department of Instrument Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering. His research interests include computer graphics and computer vision.

Zan-Wei Zhou is currently pursuing the B.S. degree in information engineering with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include 3D vision, time series forecasting, and theories of machine learning and deep learning.

Bin Ji received the B.S. degree from the Department of Mechanical Engineering, Xidian University, and the M.S. degree from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering. His research interests include computer graphics and computer vision. He is a member of the Character Laboratory, Shanghai Jiao Tong University.

Guang-Tao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. His research interests include multimedia signal processing and perceptual signal processing. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012.

Wei Shen was an Assistant Research Professor at the Department of Computer Science, Johns Hopkins University. He has been a tenure-track Associate Professor at the Artificial Intelligence Institute, Shanghai Jiao Tong University, since October 2020. His research interests lie in the fields of computer vision, machine learning, and deep learning, particularly in object detection and segmentation, representation learning, human-centered computer vision, and medical image analysis. He is the Area Chair of CVPR 2022 and ACCV 2022, a Senior Program Committee Member of AAAI 2022, and an Associate Editor of *Neurocomputing*.