# High-Fidelity Human Avatars from a Single RGB Camera

Hao Zhao[1] Jinsong Zhang[1] Yu-Kun Lai[2] Zerong Zheng[3] Yingdi Xie[4] Yebin Liu[3] Kun Li[1]*
[1]Tianjin University, China [2]Cardiff University, United Kingdom
[3]Tsinghua University, China [4]VRC Inc., Japan
{zhaohao120,jinszhang,lik}@tju.edu.cn LaiY4@cardiff.ac.uk
{zzr18,liuyebin}@mail.tsinghua.edu.cn {yingdi.xie}@vrcjp.com

## Abstract

*In this paper, we propose a coarse-to-fine framework to reconstruct a personalized high-fidelity human avatar from a monocular video. To deal with the misalignment problem caused by the changed poses and shapes in different frames, we design a dynamic surface network to recover pose-dependent surface deformations, which help to decouple the shape and texture of the person. To cope with the complexity of textures and generate photo-realistic results, we propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy to obtain detailed textures. Our framework also enables photo-realistic novel view/pose synthesis and shape editing applications. Experimental results on both the public dataset and our collected dataset demonstrate that our method outperforms the state-of-the-art methods. The code and dataset will be available at http://cic.tju.edu.cn/faculty/likun/projects/HF-Avatar.*

## 1. Introduction

Automatic generation of personalized human avatars has a wide range of applications in virtual/augmented reality, virtual try-on, entertainment and gaming. Especially technologies using a single RGB camera will enable To C (customer) applications instead of To B (business).

High-quality human models can be reconstructed with expensive 3D scanners [1], multi-view studios with controlled lighting [10], or depth cameras [6, 8, 53]. These systems are usually costly or using non-consumer devices, leading to restricted applications. Therefore, avatar acquisition from a single RGB camera is the most practical but challenging. Some methods [38, 39, 51] based on implicit representations reconstruct both geometry and texture from a single image, which can handle arbitrary topology but cannot support animation. Moreover, the reconstructed unseen regions tend to be smooth due to the limited observation. Therefore, many work proposed to reconstruct an avatar from an RGB video. Alldieck *et al*. [2–4] proposed to generalize visual hull methods to monocular videos of people in motion, which optimized a fixed displacement for each vertex across the video. Although this method is computationally cheap and memory-saving, such a single offset shared across all the frames is unreasonable, because the pose and geometry of the person change with the moving. Besides the reconstructed geometry, a high-quality texture map is also an essential component for a personalized avatar. Alldieck *et al*. [4] proposed to get a full texture map by calculating the median of unwrapped texture maps, which leads to a coarse texture map due to direct averaging. To obtain a sharp texture map, they further proposed to solve the texture stitching based on graph cut [2, 3]. However, all the above methods suffer from either blurred textures or texture artifacts/mistakes, due to the intrinsic complexity of textures and unreasonable processing (shown in

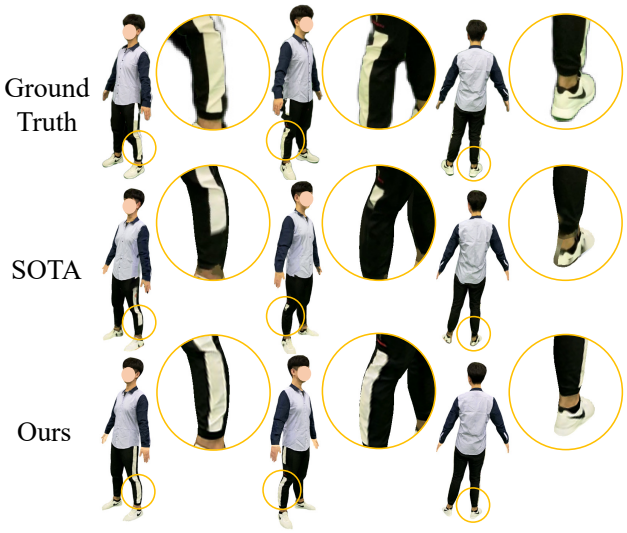


Figure 1. Given a self-captured RGB video, the state-of-the-art method [2] fails to produce seamless and reasonable texture maps. To address this, we propose a coarse-to-fine framework with dynamic surface deformation and reference-based neural rendering, which can generate seamless and sharp texture maps.

*Corresponding author

the middle of Fig. 1).

To address the above problems, in this paper, we propose a coarse-to-fine framework which consists of a dynamic surface network and a reference-based neural rendering network, to generate a fully-textured high-fidelity avatar from a monocular video where the person is rotating in front of the camera. The geometry of the person will change continuously when the person is moving. This leads to the misalignment among different frames of the video. To deal with the misalignment problem, we design a dynamic surface network to recover pose-dependent surface deformations, which help to decouple the shape and texture of the person. We learn to optimize both geometry and texture by the photometric constraint, which guides the vertices to be close to the right positions and relieves the misalignment of geometry.

Based on the dynamic surface network, we obtain a coarse texture map. However, texture is extremely complex: it resides in high dimensional space and is difficult to represent. Therefore, to cope with the complexity of textures and generate photo-realistic results, we propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy to obtain detailed textures. The neural rendering network fuses observations into a joint representation whose results are used as supervision to optimize the texture map which avoids the direct averaging of textures and adds more texture details. Besides, we propose to map the supervisions into a new space by enhancing its high-frequency information, which improves the clarity and fidelity of texture maps. Our framework can reconstruct high-fidelity personalized avatars and generate photo-realistic results of novel view/pose synthesis, which is compatible with traditional graphics pipeline. Experimental results on both the public dataset and our collected dataset demonstrate that our method outperforms the state-of-the-art methods. An example is given in Fig. 1.

The main contributions are summarized as follows:

- We propose a coarse-to-fine framework which combines neural texture with dynamic surface deformation to generate a fully-textured avatar from a monocular video captured by the users themselves.
- We propose a dynamic surface network to model the pose-dependent surface deformations of a moving person, which deals with the misalignment problem and disentangles the shape and texture of the person.
- We propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy, which fuses all the observations into a consistent representation and enables to generate the detailed texture map.

## 2. Related Work

### 2.1. Avatar Acquisition

The automatic acquisition of personalized human avatars is critical for many applications such as VR/AR, gaming, teleconferencing and virtual try-on. High-quality human models can be created with a scanner [1] or a multi-camera system [20, 40–42], but the cost and the size prevent their practical applications. Although some methods [8, 22, 48, 53] obtain high-quality 3D reconstruction by relying on depth sensors, RGB-D cameras are less ubiquitous than RGB cameras.

In order to enable the To C (to customer) applications, human reconstruction from a common RGB camera is very important. To reduce the ambiguity in monocular cases, Zheng *et al*. [52] proposed an image-guided volume-to-volume translation CNN and a dense semantic representation for human reconstruction, but they cannot recover fine-scale details. To generate detailed reconstruction, some methods [38, 39, 51] proposed to establish a pixel-aligned implicit function which can infer both geometry and texture from a single image. They learn human priors from a synthetic human dataset. Although the implicit field representations used can handle arbitrary topology, they cannot support animations. Alldieck *et al*. [5] and Lazova *et al*. [19] reconstructed a detailed parametric human model by solving an image-to-image translation problem to regress offsets in UV-space from a single RGB image, but they require a frontal photo as input and the recovered pose is restricted to A-pose. To obtain topology-consistent reconstruction for any pose, Li *et al*. [21] propose a hierarchical graph transformation network. However, the reconstructed geometry and texture from a single RGB image is still smooth for unseen parts. Alldieck *et al*. [3, 4] proposed a video-based method to transform bodies into a canonical pose and optimized the projected silhouettes, which enables efficient optimization of consistent 3D shapes. To avoid the time-consuming optimization of [3, 4], Alldieck *et al*. [2] presented a hybrid learning and optimization method, which infers a personalized avatar from a few frames of an RGB video. These video-based methods generate promising results using a single RGB camera, but cannot cope with dynamic deformation among different frames and suffer from blurred textures, stitching artifacts, or texture mistakes.

In this paper, we propose a coarse-to-fine framework to generate a fully-textured avatar from a monocular video. To deal with the inconsistent poses and shapes in different frames, we develop a novel dynamic surface network to model pose-dependent deformation, which also enables to disentangle the shape and texture of the person. To avoid texture artifacts and generate photo-realistic results, we propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy.

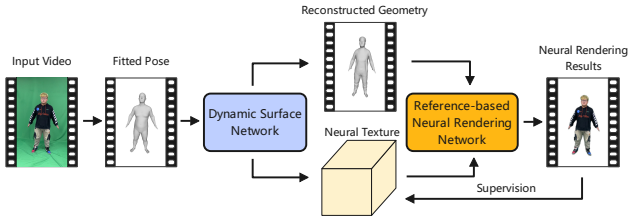


Figure 2. To generate a fully-textured avatar, we design a coarse-to-fine framework with a dynamic surface network and a reference-based neural rendering network. The dynamic surface network disentangles the shape and texture of the person and generates a coarse texture map for the initialization of neural texture. Then, we fine-tune the reference-based neural rendering network with pre-trained weights on a new subject with several epochs. After fine-tuning, we generate photo-realistic images and take them as the supervisions to optimize the texture map through back-propagation.

### 2.2. Novel View/Pose Synthesis

Neural representations and implicit fields [13, 17, 23, 24, 26, 31, 43–45] have emerged as powerful tools which can generate photo-realistic results. NeRF [30] proposed to model a scene as a continuous 5D function that maps the spatial position and viewing direction to implicit fields, which achieves high-quality rendering. Many work tried to apply NeRF on dynamic scenes by adopting a multi-camera system to get adequate information. Neural Body [33] proposed a representation where the learned latent codes are anchored to a deformable mesh to provide the network with geometric guidance. Other methods [31, 32, 34] proposed a deformation field to build the correspondence between different frames and achieved amazing results. Recently, by combing the surface field and the radiance field, DoubleField [42] achieves high-quality human reconstruction and rendering from sparse views. To better model the motion hierarchy of generic clothes, a concurrent work [50] proposed a method using a set of structured local radiance fields anchored to a human body template. However, the multi-camera system is expensive and difficult to maintain.

Recently, generative adversarial networks (GANs) [9, 14, 46] have made great progress in yielding high-fidelity images of humans. Many approaches formulate the motion transfer problem as an image-to-image translation task. Kappel *et al*. [15] divided the image translation task into four cascaded generative networks and proposed a structure network to learn wrinkles of garments, which generates high-quality results. Zhang *et al*. [49] proposed a decoupled GAN to disentangle the shape and texture of clothing. Although these methods achieve inspiring results, unnatural appearance, lost texture details and temporal inconsistency sometimes occur due to the lack of 3D information.

To reduce the above ambiguities, ANR [35] and StylePeople [12] proposed to combine a coarse parametric human model with the neural texture by extending deferred neural rendering (DNR) [45]. Although DNR is theoretically powerful, it requires an accurate geometry which is impractical in real scenes. ANR and StylePeople tried to address this problem in the neural rendering network to paint the texture outside the geometry. However, they do not disentangle the shape and texture of the person completely because they only use a coarse mesh to track the pose. Moreover, an explicit texture map is lost due to the coarse geometry, and it is also inevitable that the coarse geometry will cause artifacts.

In this paper, we reconstruct explicit high-fidelity texture maps by neural networks from monocular videos, and also achieve photo-realistic novel view/motion synthesis results. Besides, the shape and texture of the person are disentangled benefiting from our dynamic surface network.

## 3. Method

The goal of our work is to create a fully-textured high-fidelity avatar from a single RGB camera. Fig. 2 shows the framework of our method. The input is a monocular video where a person rotates with A-pose in front of the camera, and we extract the human foreground by a state-of-the-art matting method [16]. The most significant differences with existing work are that we propose a dynamic surface network to decouple the shape and texture of the person, and a reference-based neural rendering network with a novel bottom-up sharpening-guided strategy to fuse all the observations into a consistent representation to generate a seamless and sharp texture map. Our method consists of three steps: 1) dynamic surface reconstruction and coarse texture map generation (Sec. 3.1); 2) reference-based neural rendering (Sec. 3.2.1); 3) texture map refinement (Sec. 3.2.2). In order to capture the non-rigid pose-dependent deformations of the person, we design a dynamic surface function, which not only captures the pose-dependent deformation but also disentangles the shape and texture of the person. To generate a seamless and sharp texture map, we design a reference-based neural rendering network and exploit a sharpening-guided fine-tuning strategy in a coarse-to-fine manner. The neural rendering network learns a joint representation between geometry and input image, which relieves the misalignment of geometry and enables to generate sharp and seamless texture maps.

### 3.1. Geometry and Texture Map Reconstruction

Previous work [2–4] tried to reconstruct personalized geometry by extending visual hull methods to monocular cases, which cannot recover dynamic deformations. Therefore, to model the geometric deformation of the moving person and handle non-rigid deformations, we propose a dynamic surface network to predict *dynamic offsets* on the template of SMPL [27] to expand the representation capac-

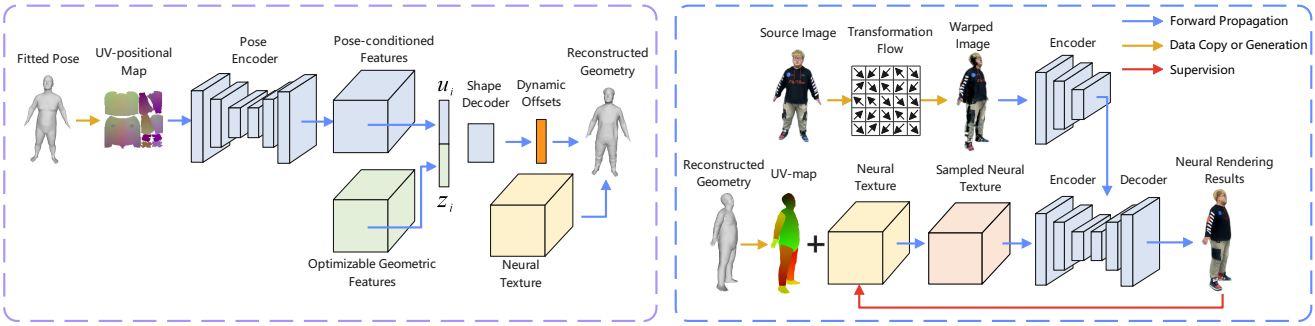


Figure 3. Left: We design a dynamic surface network which can capture pose-dependent geometric deformations. Right: We design a reference-based neural rendering network and exploit a sharpening-guided texture map generation strategy to generate seamless and sharp texture maps.

ity of SMPL:

$$
\begin{aligned}
M(\beta,\theta_i,\mathbf{D_i}) &= W(T(\beta,\theta_i,\mathbf{D_i}),J(\beta),\theta_i,\mathcal{W}),\\
T(\beta,\theta_i,\mathbf{D_i}) &= \mathbf{T}+B_s(\beta)+B_p(\theta_i)+\mathbf{D_i},\\
\mathbf{D_i} &= f_w(\theta_i),
\end{aligned}
\tag{1}
$$

where $\beta$ is the shape parameter, and $\theta_i$, $\mathbf{D_i}$ are the pose parameter and the offset vector of the $i$-th frame, respectively. $W$ is a linear blend skinning function with blend weights $\mathcal{W}$ which is applied to the morphed shape $T(\beta,\theta_i,\mathbf{D_i})$ based on the skeleton joints $J(\beta)$. The morphed shape is obtained by applying pose-dependent deformations $B_s(\beta)$, shape-dependent deformations $B_p(\theta_i)$ and dynamic offsets $\mathbf{D_i}$ to the template $\mathbf{T}$. $f_w(\cdot)$ is our dynamic surface network with weights $w$.

The details of dynamic surface network are illustrated in Fig. 3 (left). The geometric feature $z_i$ is conditioned on the pose feature $u_i$ before being the input of the shape decoder, similar to [29]. The shape decoder is represented by 8-layer multilayer perception (*MLP*) with 256 feature channels. The pose feature is encoded by a U-net [37] whose input is the UV positional map of the posed body. To make full use of the only video input, we propose to optimize both the geometry and the texture by a silhouette matching term and a photometric tracking term. Based on a differential renderer [36], the silhouette matching term penalizes the difference between the rendered silhouette using the predicted geometry and the silhouette extracted from the original image. The photometric tracking term encourages the rendered images obtained using the predicted geometry and the predicted texture to be similar to the input images. The gradients of the photometric term can be back-propagated to the vertices and guide the vertices to be close to the right positions, which further relieves the misalignment of geometry. After training the dynamic surface network, we can get an initial texture map which is used to train the neural rendering network. Given a target pose, the model outputs the 3D dynamic bodies by predicting dynamic offset fields based on the pose feature and the learned geometric feature in UV-space.

### 3.2. Detailed Texture Generation

Previous work [3, 4] generated a texture map by taking the median or selecting one out of $K$ frames, leading to blurriness and discontinuity. In particular, an effective solution needs to coherently aggregate appearance information from monocular observations across time as the body undergoes a 3D motion. However, the texture map generation method of Sec. 3.1 still cannot avoid the averaging phenomenon of textures. Therefore, based on the texture map generated in Sec. 3.1, we propose a novel texture map generation method with reference-based neural rendering and design a coarse-to-fine strategy to generate a detailed texture map, as shown in Fig. 3 (right). We fuse all the observations into a joint representation through the neural texture and the neural rendering framework. First, we learn a reference-based neural rendering network based on the input image and the reconstructed geometry to produce photo-realistic images. Then, the results of neural rendering are used as supervision to optimize the coarse texture map to have more details. Besides, to improve the sharpness and fidelity of texture maps, we propose to map supervision from the low-frequency domain to the high-frequency domain by a sharpening kernel function.

#### 3.2.1 Reference-based Neural Rendering

We obtain a relatively shape-accurate mesh by establishing an instance-specific dynamic surface function in Sec. 3.1, which disentangles the shape and texture, and makes the neural rendering network focus on texture information. Besides, we get a UV-map by adopting barycentric interpolation on the reconstructed geometry. However, only decoupling the shape and the texture is insufficient, and it is still non-trivial for the neural network to learn complex textures and patterns.

Reference-based image processing has succeeded in image super-resolution [47]. We propose a reference-based neural rendering network which transfers high-resolution textures from a given reference image, to produce photo-realistic results. We obtain an incomplete but sharp image from the input video by warping the reference image from

the current pose to that aligned with our input image using an image warping method [25]. For the sake of simplicity, we use front and back images of the person to be reconstructed for warping. With the warped image, we can transfer the texture information to the generated feature by concatenating them directly. Specifically, given a 3D geometry and a valid UV-map, we carry out bilinear sampling on neural texture and translate high-dimensional neural textures into RGB images similar to [45] using a neural network which is formulated as:

$$I = \mathcal{R}(\mathcal{T}, I_{uv}, I_{ref}), \tag{2}$$

where $I_{uv}$ is the UV-map whose pixels store the corresponding positions in the UV space, $I_{ref}$ is the warped reference image from the input images, and the neural rendering model is defined with neural texture $\mathcal{T}$ and neural rendering network $\mathcal{R}$. The number of feature channels of neural texture is 16.

#### 3.2.2 Texture Refinement

To generate a detailed texture map, we propose to use neural rendering results to optimize the coarse texture map through back-propagation. Compared with taking the input images as supervision on the texture map, the images generated by the neural rendering network are more aligned with the reconstructed geometry. In order to generate a sharp and seamless texture map, we propose to map the supervision from the low-frequency domain to the high-frequency domain through a sharpening kernel function. A sparse gradient map was proposed to guide structure-preserving image super-resolution [28]. However, most areas of the gradient map are close to zero, which cannot improve the clarity of the image. Therefore, we use an unsharp masking (*USM*) method [11] to calculate the kernel by subtracting Gaussian filter kernel from the identity kernel.

First, we map the neural rendering results to a new domain and enhance its high-frequency information by the sharpening kernel. Then, the coarse texture map is supervised by the mapped neural rendering results. We carry out bilinear interpolation on the texture map again using the UV-map where each pixel stores the corresponding position of the texture map, compute the L1 distance between the output image and the pseudo ground truth, and update the value of coarse texture map through back-propagation. After optimization, we can get a seamless and sharp texture map with an image resolution of $512 \times 512$.

### 3.3. Training Details

We first train the dynamic surface network and then train the reference-based neural rendering network. The losses are given in the supplementary document.

#### 3.3.1 Dynamic Surface Network

To stabilize the optimization process of the model, the dynamic surface network is optimized in two stages. First, we initialize with the solution of SMPLify [7] and optimize the pose, translation and shape parameters of SMPL with the supervision of detected 2D joints and silhouettes. Based on the initial parameters, we then optimize the offset and the texture jointly using *ADAM* [18]. Note that the pose-encoder is not shared across subjects, because this increases training complexity with limited improvement. After optimization, we get a dynamic geometry and a coarse texture map.

| Dataset | VideoAvatar [4] | Octopus [2] | Ours |
|---|---|---|---|
| *People-Snapshot* [4] | 39.5940 | 27.7767 | **26.4135** |
| *SelfieVideo* | 23.7101 | 16.3087 | **15.1284** |

Table 1. Quantitative comparison (FID ↓) on two datasets.

| | VideoAvatar [4] | Octopus [2] | Ours |
|---|---|---|---|
| MVE (cm) | 5.8183 | 4.5244 | **4.4547** |

Table 2. Quantitative comparison of geometry reconstruction.

#### 3.3.2 Reference-based Neural Rendering Network

Directly training the whole network will cause unstable and blurry results. Therefore, we use the generated coarse texture map as the initial value for the first three layers of neural texture and freeze it. Then, the neural texture and the neural renderer are trained end-to-end on the whole dataset. Each person has a unique neural texture, and the parameters of the neural renderer are shared. Note that, for a new person, the neural rendering network with pre-trained weights only needs to be fine-tuned with several epochs.

## 4. Experimental Results

### 4.1. Dataset

We evaluate the performance of the proposed method on *People-Snapshot* dataset [4] and our collected dataset named *SelfieVideo*. *People-Snapshot* [4] consists of 24 videos of 11 subjects, while our dataset consists of 80 videos of 80 diverse clothed persons, captured with a HD-camera with image resolution of $2160 \times 1216$. The subjects were collected from the talent market and each subject signs a license agreement. We proportionally resize the frames to $1024 \times 1024$ resolution due to memory requirement in our experiments. Each video contains about 300 frames, and all the subjects are required to rotate with A-pose in front of the camera. Please note that, although the dataset is captured with a green screen, our method is also applicable to videos with ordinary backgrounds.

### 4.2. Comparison

We compare our method on the public dataset *People-Snapshot* [4] and our dataset *SelfieVideo* with two state-of-the-art video-based avatar generation methods, VideoA-

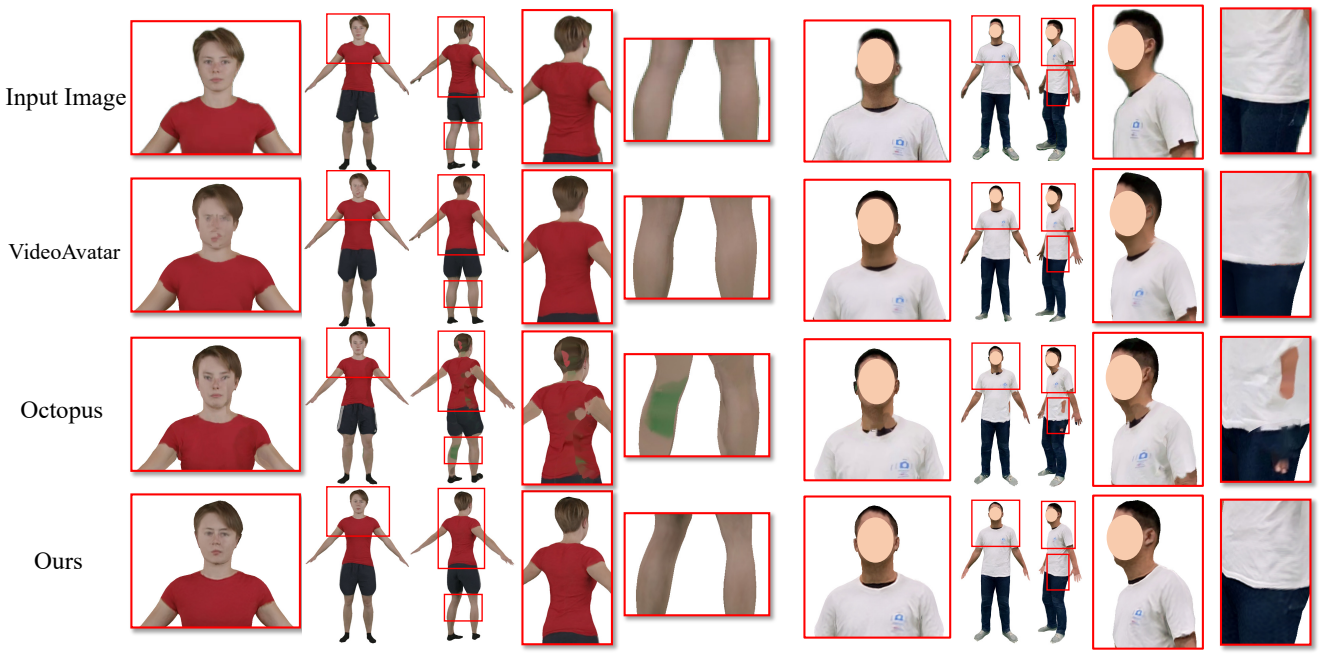


Figure 4. Reconstructed textured-avatars by VideoAvatar [4], Octopus [2] and ours on *People-Snapshot* [4] (left) and *SelfieVideo* (right).

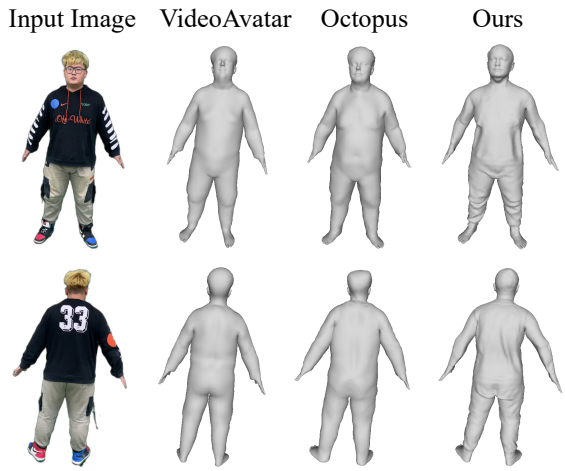


Figure 5. Reconstructed 3D geometry by VideoAvatar [4], Octopus [2] and our method.


Figure 6. Qualitative results of ablation study on different supervision and sharpening schemes.

vatar [4] and Octopus [2]. The results of the two methods are generated using the official implementations. Quantitative results on the two datasets are given in Tab. 1. Due to lack of ground truths, the existing pixel-aligned metrics, *e.g.*, PSNR, LPIPS, are not suitable. Therefore, we use FID (Fréchet Inception Distance), a metric to measure the distance between the distributions of the real images and the generated images, for our evaluation. We calculate FID between the rendered images using the generated texture map and the originally captured images. Our method achieves the best performance on both datasets, which indicates that our method generates more realistic results.

Some visual results are shown in Fig. 4. Compared with VideoAvatar [4], our method can generate sharper texture maps and clear patterns. Compared with Octopus [2], our method generates seamless texture maps, and there are no texture mistakes or lost patterns in our generated texture maps. In a word, our method can generate seamless and sharp texture map compared with the state-of-the-art methods, which benefits from our coarse-to-fine framework and sharpening-guided fine-tuning strategy. Besides, we provide qualitative comparison on the reconstructed geometry in Fig. 5. Our method can reconstruct more accurate and detailed geometry, benefiting from the design of dynamic surface network. Due to lack of ground truths, we quantitatively evaluate our method on 18 scanned human models. We adopt the same data generation protocol as [2] to register SMPL+D to each scan, and render the video by chang-

| Method | I-noS | N-noS | N-USM |
|---|---|---|---|
| FID ↓ | 33.2064 | 23.4513 | **15.1284** |

Table 3. Quantitative results of ablation study on different supervision and sharpening schemes.


Figure 7. Qualitative results of ablation study on reference branch and training scheme.

ing the pose parameters of the SMPL. The mean vertex error over the whole video across the test subjects compared with VideoAvatar [4] and Octopus [2] is given in Tab. 2. Our method achieves better results than the state-of-the-art methods [2, 4].

### 4.3. Ablation Study

**Different Supervision and Sharpening Schemes.** We study the effect of different supervision and sharpening schemes on texture map generation. Fig. 6 shows qualitative comparison of using different supervision and sharpening schemes. We compare three variants: supervised by input images without sharpening (I-noS), supervised by neural rendering results without sharpening (N-noS), and supervised by neural rendering results with unsharp masking (N-USM). From the comparison of I-noS and N-noS, we can see that the generated texture map supervised by the neural rendering results has more accurate texture details. In the last two columns, the generated texture with unsharp masking is clearer and more detailed. Quantitative results on 80 subjects in terms of FID are shown in Tab. 3. The model supervised by neural rendering results with unsharp masking also achieves the best score, which demonstrates the effectiveness of our proposed method.

**Reference Branch and Training Scheme.** Fig. 7 shows the effect of the reference branch and our training scheme for neural rendering network (Sec. 3.3.2). We compare three variants where the models are trained end-to-end without our training scheme but with the reference branch (w/o TS), trained with our training scheme but without the reference branch (w/o REF), and trained with our training scheme and the reference branch (Full). From the comparison of the last two rows of Fig. 7, the reference branch recovers more texture details and patterns. From the first row and the last row, we can see that our model can generate more reasonable and photo-realistic results with the proposed training scheme. We also design a user study for better evaluation in supplementary material.

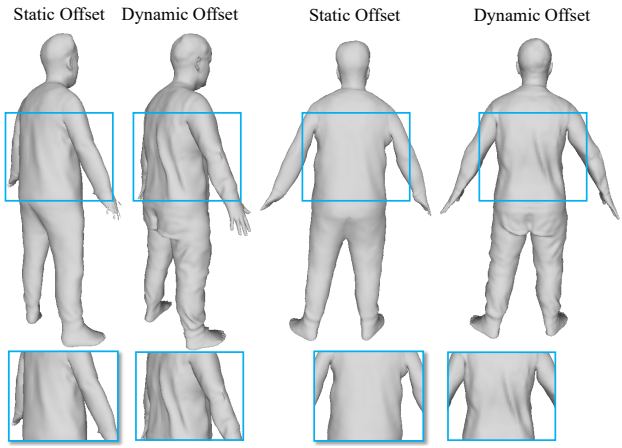


Figure 8. Comparison results of using static offset and dynamic offset. With the dynamic surface network, the geometric details can be better reconstructed compared with static offset.

| Method | [33] | [15] | [12] | Ours |
|---|---|---|---|---|
| FID ↓ | 81.8043 | 45.1285 | 63.8366 | **28.1964** |

Table 4. Quantitative comparison for novel view synthesis.

**Static Offset vs. Dynamic Offset.** To capture the pose-dependent deformation of the person from only RGB input, we design a dynamic surface network. Fig. 8 shows the comparison results of reconstructed geometries using static offset and dynamic offset. It can be seen that our dynamic surface network can recover more geometric details compared with using static offset.

### 4.4. Applications

**Novel View Synthesis.** Given a target view, we can generate a view-conditioned UV-map with rasterization using z-buffer. With the corresponding UV-map, the geometry is rasterized using a neural texture by bilinear sampling and then is translated to an RGB image using a neural network. We compare our method with three state-of-the-art methods Neural Body [33], HF-NHMT [15] and StylePeople [12]. The trained models of [33] and [15] are generated by the official implementations, and the trained models of [12] on 20 videos of *SelfieVideo* are provided by the authors. As illustrated in Fig. 9, our method achieves the most reasonable and photo-realistic results. Tab. 4 gives the quantitative results on the 20 videos. Due to lack of ground truths, FID is calculated by computing the distance between distribu-

Figure 9. Novel view synthesis results of NeuralBody [33] (top row), HF-NHMT [15] (second row), StylePeople [12] (third row), and our method (bottom row).

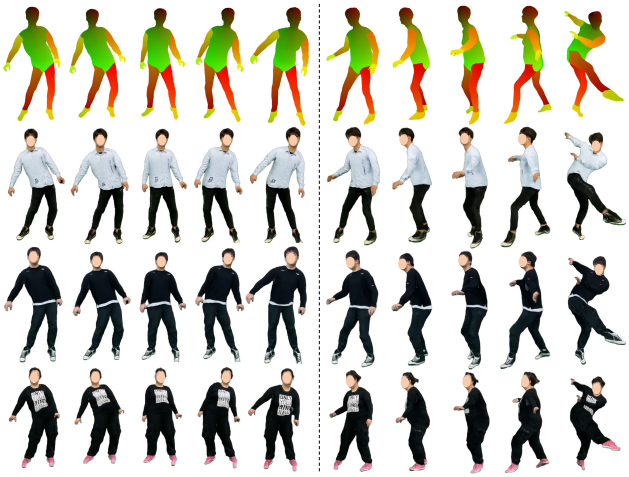

Figure 10. Novel pose synthesis results of three persons using the poses of the first row.

tions of the generated images and the captured images. Our method outperforms the other methods.

**Novel Pose Synthesis.** Given a target pose, we can also generate a pose-conditioned UV-map with rasterization using z-buffer. The learned person can be retargeted to the poses from the pre-captured motion sequences. Fig. 10 shows the generated results of different persons with the same poses.

**Shape Editing.** Benefiting from our design of the dynamic surface network which disentangles the shape and texture of the person, our method can achieve shape editing by changing the parameters of the SMPL model. Fig. 11 shows some neural rendering results of one person with the upper-bodies

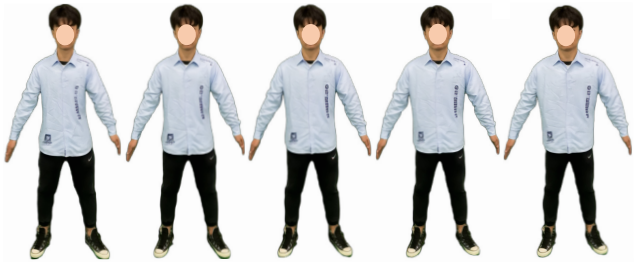

Figure 11. The results of shape editing. From left to right, we show the results of person changing from thin to fat.

changing from thin to fat. It can be seen that the texture is not distorted as the shape changes, which proves that our method can disentangle the shape and texture of the person.

## 5. Conclusion and Discussion

**Conclusion.** In this paper, we propose a novel framework to generate a fully-textured avatar and demonstrate high-fidelity rendering for novel view/pose synthesis. We introduce the dynamic surface network to capture the pose-dependent deformations of the person and the reference-based neural rendering network to generate high-fidelity images for view/pose synthesis, and we exploit a novel sharpening-guided strategy to generate seamless and sharp texture maps. We study the effect of our proposed module in ablation study, and our method is a step forward in avatar generation and neural rendering.

**Limitations.** Although we have achieved high-fidelity avatar generation from a single RGB camera, there are still some cases that we cannot solve well: very loose clothes, *e.g.*, dresses, wrong pose estimation caused by depth ambiguity, dynamic appearance effects, *e.g.*, pose-dependent wrinkles, and extremely complex textures. In further work, we will combine efficient implicit representations, *e.g.*, implicit surfaces and NeRFs, to break through the limitation of the fixed topology, improve the representation capacity of the framework and generate more high-fidelity avatars.

**Broader Impact.** Similar to DeepFake (FaceSwap), our method can make anyone have the possibility to participate in avatar generation and AI development, which may cause private and ethical problems. Therefore, we propose that the reconstructed avatar should be encrypted into a special format and it can only be decoded by the specific software, and the people who use the technology must sign an agreement and promise not to harm the privacy of the public. Besides, we suggest policymakers to establish an efficient regulatory system to avoid the disclosure of personal information. We believe that the development of the technology will promote the improvement of related policies.

**Acknowledgments.** We thank Artur Grigorev (the author of StylePeople [12]) for the provided pre-trained models. This work was supported in part by the National Natural Science Foundation of China (62171317, 62125107 and 62122058).

## References


[1] https://www.artec3d.cn/. 1, 2

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 3, 5, 6, 7

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, 2018. 1, 2, 3, 4

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 3, 4, 5, 6, 7

[5] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *Int. Conf. Comput. Vis.*, 2019. 2

[6] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Int. Conf. Comput. Vis.*, 2015. 1

[7] 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 *Eur. Conf. Comput. Vis.*, 2016. 5

[8] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Int. Conf. Comput. Vis.*, 2021. 1, 2

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *arXiv:1808.07371*, 2018. 3

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4):1–13, 2015. 1

[11] Neycenssac F. Contrast enhancement using the laplacian-of-a-gaussian filter. *Graphical Models and Image Processing*, 55(6):447–463, 1993. 5

[12] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. StylePeople: A generative model of fullbody human avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3, 7, 8

[13] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4), 2021. 3

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[15] Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus Magnor. High-fidelity neural human motion transfer from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3, 7, 8

[16] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson W.H. Lau. Is a green screen really necessary for real-time portrait matting? *arXiv:2011.11961*, 2020. 3

[17] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[19] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-Degree textures of people in clothing from a single image. *arXiv:1908.07117*, 2019. 2

[20] Kun Li, Qionghai Dai, and Wenli Xu. Markerless shape and motion capture from multiview video sequences. *IEEE Trans. Circuits Syst. Video Technol.*, 21(3):320–334, 2011. 2

[21] Kun Li, Hao Wen, Qiao Feng, Yuxiang Zhang, Xiongzheng Li, Jing Huang, Cunkuan Yuan, Yu-Kun Lai, and Yebin Liu. Image-guided human reconstruction via multi-scale graph transformation networks. *IEEE Trans. Image Process.*, 30:5239–5251, 2021. 2

[22] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3D self-portraits in seconds. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Adv. Neural Inform. Process. Syst.*, 2020. 3

[24] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *arXiv:2105.01794*, 2021. 3

[25] Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid Warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Int. Conf. Comput. Vis.*, 2019. 5

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural Volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, 2019. 3

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015. 3

[28] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 5

[29] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Int. Conf. Comput. Vis.*, 2021. 4

[30] 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. In *Eur. Conf. Comput. Vis.*, 2020. 3

[31] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Int. Conf. Comput. Vis.*, 2021. 3

[32] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. In *Int. Conf. Comput. Vis.*, 2021. 3

[33] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3, 7, 8

[34] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *arXiv:2011.13961*, 2020. 3

[35] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. ANR-Articulated neural rendering for virtual avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[36] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501*, 2020. 4

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597*, 2015. 4

[38] 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 *Int. Conf. Comput. Vis.*, 2019. 1, 2

[39] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Eur. Conf. Comput. Vis.*, 2016. 2

[42] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3

[43] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning persistent 3D feature embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 3

[44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene Representation Networks: Continuous 3D-structure-aware neural scene representations. *arXiv:1906.01618*, 2019. 3

[45] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred Neural Rendering: Image synthesis using neural textures. *arXiv:1904.12356*, 2019. 3, 5

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[47] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 4

[48] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[49] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. PISE: Person image synthesis and editing with decoupled gan. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[50] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[51] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 1, 2

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *Int. Conf. Comput. Vis.*, 2019. 2

[53] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G. Narasimhan, and Minh Vo. TexMesh: Reconstructing detailed human texture and geometry from RGB-D video. In *Eur. Conf. Comput. Vis.*, 2020. 1, 2