R^{2}Human: Real-Time 3D Human Appearance Rendering from a Single Image

Yuanwang Yang[1,†], Qiao Feng[1,†], Yu-Kun Lai[2], Kun Li[1,*]

[1]Tianjin University, China [2]Cardiff University, United Kingdom

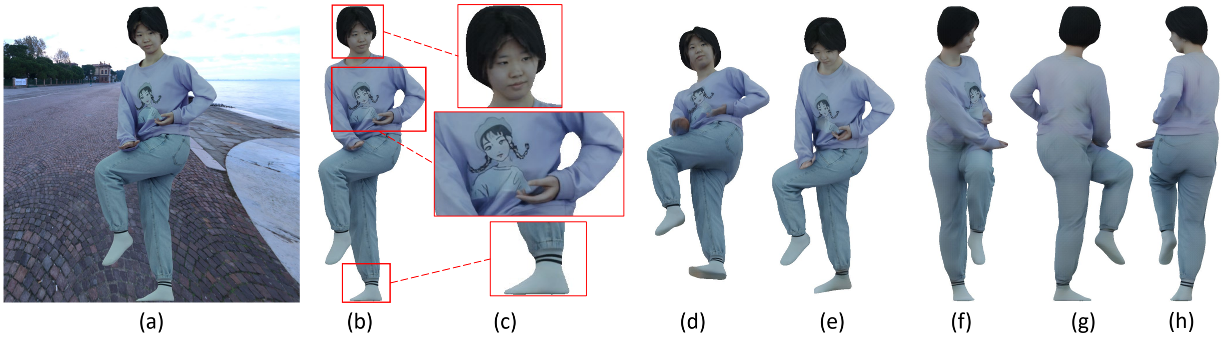

Figure 1: Given a single RGB image, R^{2}Human can generate photorealistic 3D human appearance in real-time. We utilized the input (a) to render close views (b) and zoomed in (c). The results of pitch angle changes are shown in (d-e), and (f-h) demonstrate the outcomes of divergent views.

Abstract

Rendering 3D human appearance from a single image in real-time is crucial for achieving holographic communication and immersive VR/AR. Existing methods either rely on multi-camera setups or are constrained to offline operations. In this paper, we propose R^{2}Human, the first approach for real-time inference and rendering of photorealistic 3D human appearance from a single image. The core of our approach is to combine the strengths of implicit texture fields and explicit neural rendering with our novel representation, namely Z-map. Based on this, we present an end-to-end network that performs high-fidelity color reconstruction of visible areas and provides reliable color inference for occluded regions. To further enhance the 3D perception ability of our network, we leverage the Fourier occupancy field as a prior for generating the texture field and providing a sampling surface in the rendering stage. We also propose a consistency loss and a spatial fusion strategy to ensure the multi-view coherence. Experimental results show that our method outperforms the state-of-the-art methods on both synthetic data and challenging real-world images, in real-time. The project page can be found at http://cic.tju.edu.cn/faculty/likun/projects/R2Human.

Index Terms: 3D human appearance, rendering, single image, real-time

1 Introduction

Rendering 3D human appearance from a single image in real-time not only enhances visual experiences but also paves the way for practical applications in AR/VR, enabling users to engage in immersive experiences with timely feedback. However, existing methods heavily rely on multi-camera setups [40, 23, 19, 24] or are constrained to offline operations [2, 9, 20, 1], leaving real-time 3D human rendering from a single image as an unresolved challenge. In this paper, we aim to translate a 2D human image into a vibrant 3D appearance in real-time, facilitating holographic communication, and enhancing user experience and immersion in VR/AR.

Research in human novel view synthesis has explored various methodologies, which can be roughly classified into two categories. The first category of methods [3, 23, 40] is founded on flow-based neural rendering, which warps the feature map from the input images to the target view and then produces high-quality novel view results with convolutional neural networks (CNNs). However, the network's predictions for occluded regions in the input image usually exhibit significant randomness due to the lack of 3D structure understanding, making it challenging to obtain consistently accurate results. To deal with this, these methods require multi-view input images to cover the rendering area as comprehensively as possible. Consequently, generating photorealistic renderings from only one image becomes difficult for them.

Another category of methods aims to recover a 3D consistent appearance for the human body. PIFu [20] recovers the 3D human geometry and appearance from a single human image using an occupancy field and a texture field, respectively. However, the rendered images are of low quality and the process is time-consuming. Its follow-up work [2] disentangles lighting from texture and redesigns the loss functions, resulting in improved visual quality. On the other hand, works utilizing the Neural Radiance Fields (NeRF) [17, 19, 24, 9] predict the density and radiance throughout the 3D space for rendering images by integrating along camera rays. However, NeRF-based methods often struggle with single-view inputs. While approaches above vary in human appearance representations, they share a common paradigm: they estimate geometry and color for discrete sampled points independently. This makes it difficult for the models to discern relationships between

[†] Equal contribution.
[*] Corresponding author.

the sampled points, resulting in relatively lower quality in the synthesized novel view images. Additionally, these methods require dense spatial sampling, which results in high computational costs and makes it challenging to balance real-time performance with rendering quality.

In this paper, we introduce R^2*Human*, a real-time framework for rendering human appearance from a single image, which uniquely combines flow-based rendering techniques with an implicit 3D human geometry representation to synthesize novel view images. Our method only requires a single image as input and achieves both high-quality rendering performance, as shown in Fig.1. Unlike previous human avatar animation typically involves creating actions beyond real-world scenarios, our focus lies in enhancing existing visual experiences by transforming real 2D human images into immersive 3D appearances in real-time. By converting existing 2D images to 3D, we cleverly avoids the increase in computational consumption caused by animation rendering and the decrease in the authenticity of the results. Moreover, our method can generate novel view renderings of humans in various clothing without individual-specific training. We introduce an intermediate representation, namely the Z-map, during the rendering process, which collects the source view depth of the rendered points of the target view and forms them to a 2D map. It helps lift the 2D image feature into a 3D texture feature field, while maintaining compatibility with 2D neural rendering networks. This unique capability enables our network to learn data-driven 2D appearance knowledge of clothed human images, resulting in more accurate renderings. By preserving the features of occluded points and leveraging the Z-map, our method effectively resolves depth ambiguities. Additionally, we employ an efficient 3D object representation known as Fourier occupancy fields (FOF) [5], which explicitly represents a 3D object as a multi-channel image. It can serve both as a prior for texture field generation and as a sampling surface during the rendering stage, avoiding the high computational costs caused by dense sampling. In order to ensure the consistency between multiple views and reduce the jittering phenomenon between neighboring views, we propose a consistency loss to regularize this process and design a spatial fusion strategy to enhance this in practical applications. Our method paves the way for the practical applications of AR/VR, which can be applied in holographic communication in the future. *Source code will be available for research purposes.*

In summary, the contributions of our work are as follows:

- We present a novel system for high-quality, real-time synthesis of human novel view images with only a single RGB image input. To the best of our knowledge, this is the first system to restore the full-body appearance of a 3D human in real-time from a single image, paving the way for practical applications in AR/VR.
- We propose R^2*Human*, an end-to-end CNN-based neural rendering method that combines the strengths of implicit texture field and explicit neural rendering, which can produce results with both 3D consistency and high visual quality.
- We introduce an intermediate representation called Z-map, which alleviates depth ambiguities in rendering, enabling high-fidelity color reconstruction for the visible area while providing reliable color inference for the occluded regions.
- We propose a consistency loss to ensure the multi-view coherence and reduce the jittering phenomenon, and design a spatial fusion strategy to enhance this in practical applications.

Table 1: Comparison with state-of-the-art methods. ✗: not supported, ✓: supported.

Method	Monocular Input	Real-time Processing	Fully-body Rendering	High-quality Output
Project Starline [14]	✗	✓	✗	✓
3DTexture [29]	✗	✗	✓	✗
Floren [23]	✗	✗	✓	✓
HDHuman [40]	✗	✗	✓	✓
PIFu [20]	✓	✗	✓	✗
SHERF [9]	✓	✗	✓	✗
Ours	✓	✓	✓	✓

2 Related work

2.1 Monocular 3D Human Reconstruction

Human reconstruction has long been a concern in the domain of computer vision and graphics. Methods for predicting human bodies from a single image represent 3D shapes primarily by estimating the properties of points in 3D space. Some previous works [39, 27] directly predict the occupancy field of a given segment in space through regression networks, but such methods have high memory requirements, which limits the spatial resolution of shape estimation. Another type of methods involves using implicit function networks to remove resolution constraints. Saito *et al*. [20] proposed an implicit function based on pixel-aligned features to reconstruct a 3D human from a single image. PIFuHD [21] further introduces normal maps to improve geometric details. Subsequent methods [38, 33, 32, 36] improve the robustness of the results by incorporating parametric SMPL [15] priors. Albahar *et al*. [1] combine diffusion model to obtain detailed geometry and texture, but the reconstruction effect of loose clothing is not good. Some methods [41, 7, 10, 31, 12, 11] take advantage of the features of SMPL to align different poses to obtain detailed reconstructions from video, but cannot adapt to single image settings. Although these methods can achieve realistic results, they often require significant computational resources and are time-consuming. Feng *et al*. [5, 6] proposed an efficient 3D geometric representation called Fourier Occupancy Field, which can establish a strong link between 3D geometry and 2D images, significantly reducing the computational requirements for reconstruction. However, one limitation is its inability to estimate texture during the geometry reconstruction process. Other work [26, 16] relies on RGBD cameras to provide depth information, thus ensuring real-time performance and reconstruction accuracy. But their results rely too much on the SMPL model, making it impossible to adapt to a variety of clothes, and the use of RGBD cameras also limits its application in daily life.

2.2 Human Novel View Synthesis

Flow-based neural rendering methods can synthesize realistic novel human view images. Shao *et al*. [23] estimate robust appearance flow with epipolar constraints, reduce ambiguous texture warping, and then synthesize high-quality novel human perspective. Zhou *et al*. [40] first estimate highly detailed 3D human geometry, and then effectively solve the serious occlusion problem caused by sparse views through geometry-guided pixel-wise feature integration method. Although these methods reduce the impact of occlusion on flow-based texture warping, they still require multi-view information to ensure the robustness of the results. Phong *et al*. [18] warp images from a single sparse RGB-D input by sphere-based rendering and refine the resulting image using an additional occlusion-free input, but its application is limited by its reliance on depth sensors. On the other hand, some methods [20, 33, 34] that focus on geometry can predict textures during reconstruction, and then synthesize novel views through traditional rendering methods.

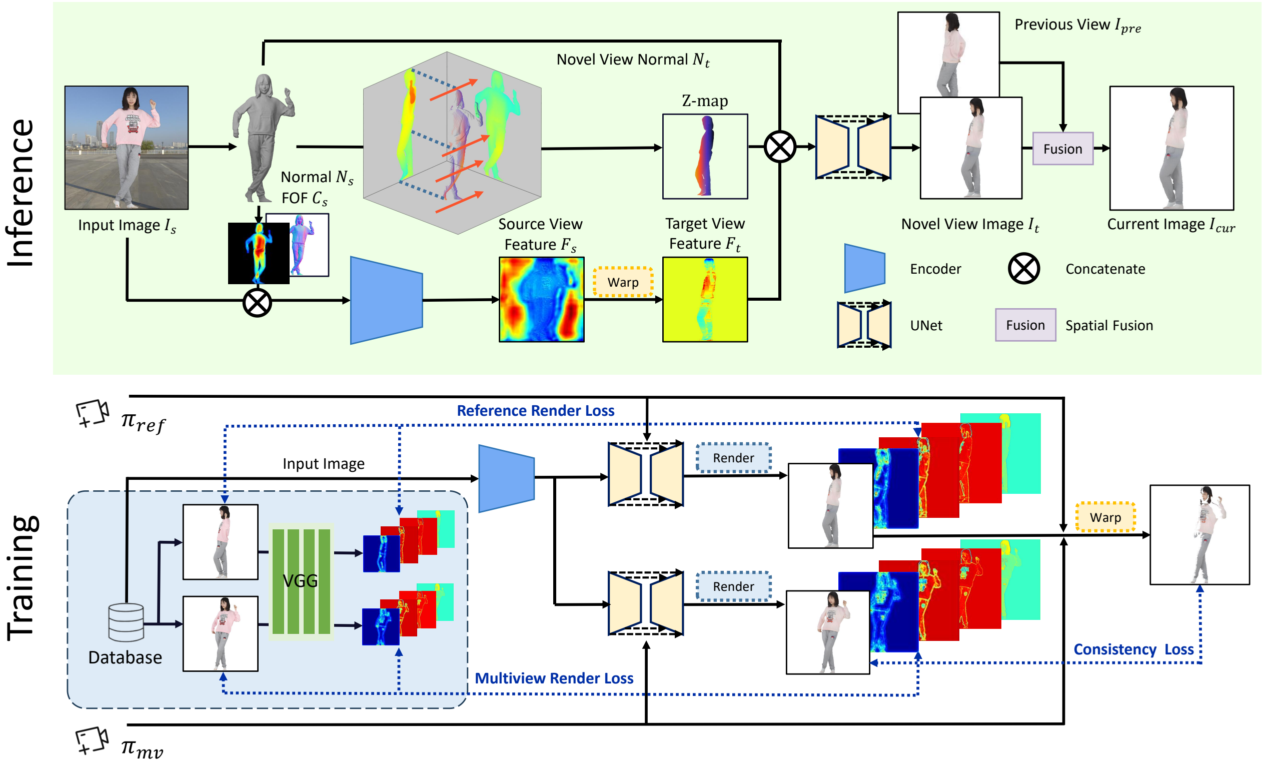

Figure 2: The overall pipeline of R^2*Human* for real-time 3D human appearance rendering. R^2*Human* leverages the proposed Z-map to combine the strengths of implicit texture field and explicit neural rendering seamlessly. With our consistency loss, we constrain the rendered color of the same visible point to be consistent across different views, thereby ensuring the multi-view consistency of the results.

Thiemo *et al*. [2] improved the visual fidelity of the results by computing albedo and shading information and carefully designing rendering losses. However, these methods find it difficult to learn the relationships between points in space due to the use of discrete sampling points to estimate color, resulting in relatively poor rendered image quality.

In recent years, neural radiance fields (NeRF) [17] have become a powerful novel view synthesis tool, using multi-layer perceptrons (MLP) to model a scene as density and color fields, resulting in realistic rendering. Researchers have made many attempts on the basis of NeRF. Peng *et al*. [19] anchor the learned potential encoding onto SMPL [15] fixed points to integrate multi-view video information, but it can only render humans with relatively uniform textures. Shao *et al*. [24] combine the surface field and the radiance field to achieve high-quality rendering, but are still limited by the need for multiple views. Hu *et al*. [9] proposed the first generalizable Human NeRF based on a single image input, but visible artifacts still exist in partially occluded bodies in the observation space. In addition, these implicit representation based methods all have a common problem: due to the high computational complexity required for dense sampling, they are time-consuming.

In this paper, we contribute the first real-time and high-quality human novel view synthesis system from a single RGB image, which combines the strengths of implicit texture field and explicit neural rendering and translates a 2D human image into a vibrant 3D appearance in real-time, facilitating holographic communication, and enhancing user experience and immersion in VR/AR. Tab. 1 presents a comparative analysis focusing on key features: support for monocular input, real-time processing capabilities, full-body rendering, and high-fidelity novel view synthesis. As indicated in the table, our method is unique in its support for all these features, distinguishing it from other approaches.

3 Method

Our goal is to build an end-to-end trainable framework that can achieve photorealistic novel view rendering of humans from a single RGB image. We employ FOF [5] to represent the 3D human form as a multi-channel image, thereby reducing computational time and memory overhead. Additionally, we introduce Z-map to effectively address the issue of prediction ambiguity for occluded points within the input image. Consequently, our network is capable of producing high-quality renderings of novel views using just a single image.

As shown in Fig. 2, our framework mainly consists of two parts: 1) We develop an image encoder to generate a texture field aligned with the reconstructed 3D geometry. 2) We warp the texture field and obtain other priors, and then use a rendering network to synthesize results with high visual quality. At inference, we estimate a detailed mesh from the input image, and using a pixel-aligned feature encoder to extract texture features. Next we mitigate depth blur in rendering with the help of Z-map representations, resulting in high-quality textures. During training, we use the input image to render two supervised views. We warp the first image to the second and perform visibility inference. By supervising the color of the visible points, we ensure the multi-view consistency of the results.

3.1 Z-map

3.1.1 Definition of Z-map

We propose a new representation called Z-map, which combines the strengths of flow-based rendering with implicit-field-based rendering. Unlike existing approaches that utilize the depth-map of the novel view for appearance rendering, our Z-map comprises the source view z-coordinate of the visible points in the novel view, aligned to the novel view. By leveraging the source view image features and the z-coordinate, the Z-map uniquely determines the position of visible points in the implicit field, thus eliminating depth ambiguity.

Most existing methods taking a single image as input rely on a pixel-aligned implicit function to produce a 3D field, such as a texture field. The color of a point in 3D space can be predicted by:

$$c = f_{mlp}(e, z), \tag{1}$$

where $c \in \mathbb{R}^3$ is the predicted RGB color, e is the corresponding feature sampled from the image feature map, z is the source view depth of that point, and f_{mlp} is the MLP-based decoder. In the task of view synthesis, each pixel on the output image I corresponds to a point in the 3D space. Thus, we can collect the colors c for all those points to generate the final result.

Such a schema does not consider the internal relationship between those points. In contrast, flow-based methods decode those features simultaneously to avoid such a drawback, which can be written as:

$$I = f_{cnn}(F_{map}(\mathscr{F}_s)), \tag{2}$$

where $\mathscr{F}_s$ is the feature map of the source view, F_{map} is the wrapping flow, and f_{cnn} is the CNN-based decoder.

Let $\mathscr{F}_t = F_{map}(\mathscr{F}_s)$ be the feature map corresponding to the target view. We can notice that $\mathscr{F}_t$ is the collection of e in Eq. 1. Therefore, if we collect all those source view depth values and form them into a 2D Z-map, these two kinds of methods can be unified in a single framework:

$$I = f_{cnn}(\mathscr{F}_t, Z_{map}), \tag{3}$$

where Z_{map} is the Z-map. Like the warping flow, Z-map can also be directly produced, which is more efficient than calculating z values for each point separately.

3.1.2 Calculation of Z-map

To combine the strengths of implicit texture field and explicit neural rendering, we propose the Z-map, which enables rendering networks to achieve 3D perception and avoid one-to-many problems caused by depth ambiguity. We define Z-map as the depth obtained from the transformation of visible points in the novel view to the source view, so we can obtain Z-map while calculating the flow maps. Specifically, we render the depth maps D_s, D_t first for both views using the predicted mesh based on camera parameters. Then, for each coordinate point $v_t \in \mathbb{R}^2$ in the novel view, we inverse project it to the world coordinate system using the camera parameters, and then project it to the source view space to obtain flow maps and Z-map from the novel view to the source view:

$$(F_{map}, Z_{map}) = \Pi^s((\Pi^t)^{-1}(v_t, D_t)), \tag{4}$$

where Π^s is the projection matrix transforming points from the world coordinates to the source view coordinates v_s and $(\Pi^t)^{-1}$ is the matrix transforming points from the novel view coordinates v_t to the world coordinates.

3.2 R^{2}Human Networks

3.2.1 Pixel-aligned Feature Encoder

To leverage the geometric information obtained from the reconstruction network and enhance feature extraction, our encoder incorporates additional information derived from the reconstructed geometry. This allows for a more comprehensive understanding of the input image. This approach consists of two key components: FOF and normal map.

FOF. As shown in Fig. 2, we obtain FOF through the existing method. It is a multi-channel image that contains 3D information and enables the encoder to capture and represent spatial relationships, depth cues, and other geometric aspects inherent in the Human. This can lead to a more robust and information-rich representation of features, potentially improving the performance of subsequent rendering.

Normal map. We generate a normal map by calculating the surface normals for each point in the predicted mesh. It can enhance the encoder's ability to perceive lighting information and geometric details.

In summary, our encoder network E is defined as follows:

$$\mathscr{F}_s = E(\oplus(I_s, C_s, N_s)), \tag{5}$$

where $\oplus$ is the concatenation operation. I_s, C_s, N_s are the input color image, the predicted FOF and estimated surface normal map, respectively.

3.2.2 Novel View Rendering

Similar to the encoder, our rendering network also integrates some additional information to improve rendering performance. Note that we do not integrate Fourier occupation fields because rendering only focuses on the visible part of the image. We use novel camera parameters Π^t to render the normal map. Then we warp the features of the source view to the novel view based on the flow map:

$$\mathscr{F}_t(v) = \mathscr{F}_s(F_{map}(v)), \tag{6}$$

In summary, our rendering network R is defined as follows:

$$I_t = R(\oplus(\mathscr{F}_t, Z_{map}, N_t)), \tag{7}$$

where N_t is the normal map rendered by novel camera Π^t.

3.2.3 Spatial Fusion Strategy

To maintain consistency during the rotation of the view, we introduce a spatial fusion strategy for generating the free-view video stream. Specifically, during the rotation process, we keep the image of the previous view I_{pre}, and use Eq. 9 to get the surface points v that are visible in both the current and previous views. Subsequently, the color of v in the current view image I_{cur} is interpolated as:

$$I_{cur}(v) = \alpha I_t(v) + (1-\alpha) I_{pre}(v), \tag{8}$$

where $\alpha = \beta |sin(\phi/2)|$, ϕ is the rotation angle of the current view relative to the input view. The closer the current view is to the input view, the more we prefer the color of I_t, while the farther the current view is from the input view, the more I_{pre} we use to ensure consistency. We set β to 0.8 in our experiments.

3.3 Training

We use synthetic data to train R^{2}Human during the training stage. As illustrated in Fig. 2, an input image along with its corresponding FOF and normal map, sampled from the training set, serves as the input. The encoder then generates the corresponding texture field features, denoted as $\mathscr{F}$. With the camera parameters Π (comprising camera direction and position), we compute the warping flow which contains the novel view information we need, enabling the generation of an image of any view with the decoder. In each gradient step, we synthesize two images $\hat{I}_{ref}$, $\hat{I}_{mv}$ with the same texture field features from a reference camera Π^{ref} and an additional camera Π^{mv} to perform multi-view consistency supervision.

Consistency Loss. To guarantee consistency between multiple views and reduce the jittering phenomenon during the rotation of the view, we propose a consistency loss to supervise the multi-view images rendered from the same texture field features. Specifically, we use Eq. 4 to calculate the flow map between the reference camera Π_{ref} and the multi-view camera Π_{mv}. Then we warp the image $\hat{I}_{ref}$ rendered by the reference camera to the muti-view camera view to get $\hat{I}^{mv}_{ref}$. Visibility inference is then performed:

If the difference between the projection depth of reference view $D^{mv}_{ref} = \Pi^{mv}((\Pi^{ref})^{-1}(v, D_{ref}))$ and the multi-view depth D_{mv} is lower than a threshold, we treat coordinate point v as visible:

$$|D^{mv}_{ref}(v) - D_{mv}(v)| < \lambda min(D^{mv}_{ref}(v), D_{mv}(v)), \tag{9}$$

where λ is a hyper-parameter and we set it to 0.02 in all our experiments.

We use the L_1 loss to supervise the visible points in the reference camera and the multi-view camera. The loss function $L_{consistency}$ can be expressed as:

$$L_{consistency} = \frac{1}{|M^{mv}_{ref}|} \sum_{(x,y)\in M^{mv}_{ref}} \|\hat{I}^{mv}_{ref}(x,y) - \hat{I}_{mv}(x,y)\|_1, \tag{10}$$

where M^{mv}_{ref} is the set of visible points between reference camera and multi-view camera calculated by Eq. 9.

Pixel Loss. Let the ground truth of the reference view and multi-view be I_{ref} and I_{mv}. We add an L_1 loss to supervise the pixel points between both sets of pairs $(I_{ref}, \hat{I}_{ref})$ and $(I_{mv}, \hat{I}_{mv})$, and to make the network more focused on the human, we only supervise the human foreground region of the image:

$$L_{pixel} = \frac{1}{|M|} \sum_{(x,y)\in M} \|I(x,y) - \hat{I}(x,y)\|_1, \tag{11}$$

where M is the set of foreground pixels of I.

LPIPS Loss. In addition, to enhance the visual effect of the output image, we also add the LPIPS loss described in [37] for both sets of pairs $(I_{ref}, \hat{I}_{ref})$ and $(I_{mv}, \hat{I}_{mv})$:

$$L_{LPIPS} = \frac{1}{|M|} \sum_{(x,y)\in M} \|VGG(I(x,y)) - VGG(\hat{I}(x,y))\|_2, \tag{12}$$

where VGG is the network described in [25].

Finally, our loss can be formulated as follows:

$$L = \lambda_1 L_{consistency} + \lambda_2 L_{pixel} + \lambda_3 L_{LPIPS}, \tag{13}$$

where we set λ_1, λ_2 and λ_3 to 100, 1 and 0.5 in our experiments.

Implement Details. Our our network is trained on synthetic data including pairs of meshes and rendered images. We collect 526 high-quality human scans from THuman2.0 dataset [35] with a wide range of clothing, poses and shapes. We randomly split them into a training set of 368 scans and a testing set of 105 scans. The remaining subjects are used as the validation set. For training our network, instead of using FOF constructed from the ground truth geometry, we apply the FOF-SMPL reconstruction network [5] to predict the FOF representation from the single-view input image. Using the predicted FOF for training can make the rendering network obtain stronger generalization ability and improved robustness. We implement our method using PyTorch and train all network components jointly, end-to-end, using the Adam optimizer, with a learning-rate of 2×10^{-4}. We train the network for 10 epochs which takes about 4 days with a batch size of 4 using a single NVidia RTX3090 GPU.

4 Experiments

4.1 Comparisons

Metrics. We use three widely used metrics for images to quantitatively evaluate our method: structural similarity (SSIM) [30], peak signal-to-noise ratio (PSNR) [22], and learned perceptual image patch similarity (LPIPS) [37] that uses AlexNet [13] to extract features.

Baselines. We quantitatively compare our approach with two methods [20, 9]: 1) PIFu [20] uses implicit functions based on pixel-aligned features to reconstruct 3D humans from a single image, and we use a standard rendering pipeline to render it to obtain a novel view image. 2) SHERF [9] is the first generalizable Human NeRF model to recover 3D humans from a single image, achieving state-of-the-art performance compared with previous generalizable Human NeRF methods. Additionally, to showcase the necessity of utilizing neural rendering, we conducted qualitative comparisons with state-of-the-art texture mapping methods 3DTexture [29].

Qualitative results. Fig. 3 and Fig. 13 presents qualitative comparisons of our approach to the baseline on the 2k2k dataset [8] and the THuman2.0 dataset [35], respectively. PIFu demonstrates the ability to estimate reasonably accurate colors, but its performance is hindered by limitations in geometric estimation, resulting in suboptimal image quality. On the other hand, while SHERF can produce results with accurate pose, it encounters challenges in generating desirable outcomes for individuals wearing loose clothing, mainly due to its heavy reliance on SMPL priors. Additionally, since SHERF estimates colors for sampling points independently, it does not effectively consider the interrelationship between these points, often leading to ambiguous results. Fig. 4 shows a qualitative comparison with traditional texture mapping methods. 3DTexture has a noticeable concatenation artifacts in the multi-view settings and cannot render occluded areas in the input view. Thanks to Z-map, R^{2}human is able to incorporate geometric priors to enhance the rendering of occluded areas by 3D information. Furthermore, with the aid of neural rendering, R^{2}human can effectively compensate for any inaccuracies in the geometric estimates. As a result, R^{2}human is capable of synthesizing high-fidelity results that exhibit remarkable 3D consistency.

Quantitative evaluations. We selected 105 models in the 2k2k and THuman2.0 test sets for evaluation, respectively. Tab. 2 shows the performance of the method when the novel view is in close proximity to the source view (with a difference of $\leq 45°$), and Tab. 3 shows the performance when the novel view is significantly different from the source view (with a difference of $\geq 90°$). It is evident that R^{2}Human outperforms SHERF and PIFu across all evaluation metrics, regardless of whether there are minor or major changes in the viewing angle.

Table 2: Quantitative rendering results in close views.

Method	THuman2.0			2k2k		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
PIFu [20]	0.8576	19.45	0.1413	0.8648	20.40	0.1139
SHERF [9]	0.8972	23.15	0.1078	0.8864	22.23	0.1113
Ours	**0.9415**	**26.92**	**0.0478**	**0.9134**	**24.69**	**0.0656**

Table 3: Quantitative rendering results in divergent views.

Method	THuman2.0			2k2k		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
PIFu [20]	0.8409	18.68	0.1681	0.8441	19.14	0.1447
SHERF [9]	0.8808	21.68	0.1249	0.8708	20.91	0.1283
Ours	**0.9221**	**25.36**	**0.0660**	**0.8889**	**22.75**	**0.0914**

Comparison of running times. Tab. 4 shows the comparison results in terms of running time. We use TensorRT to accelerate inference in the real-time system. It can be seen that our method runs significantly faster than the other two baseline methods, with a speed improvement of two orders of magnitude.

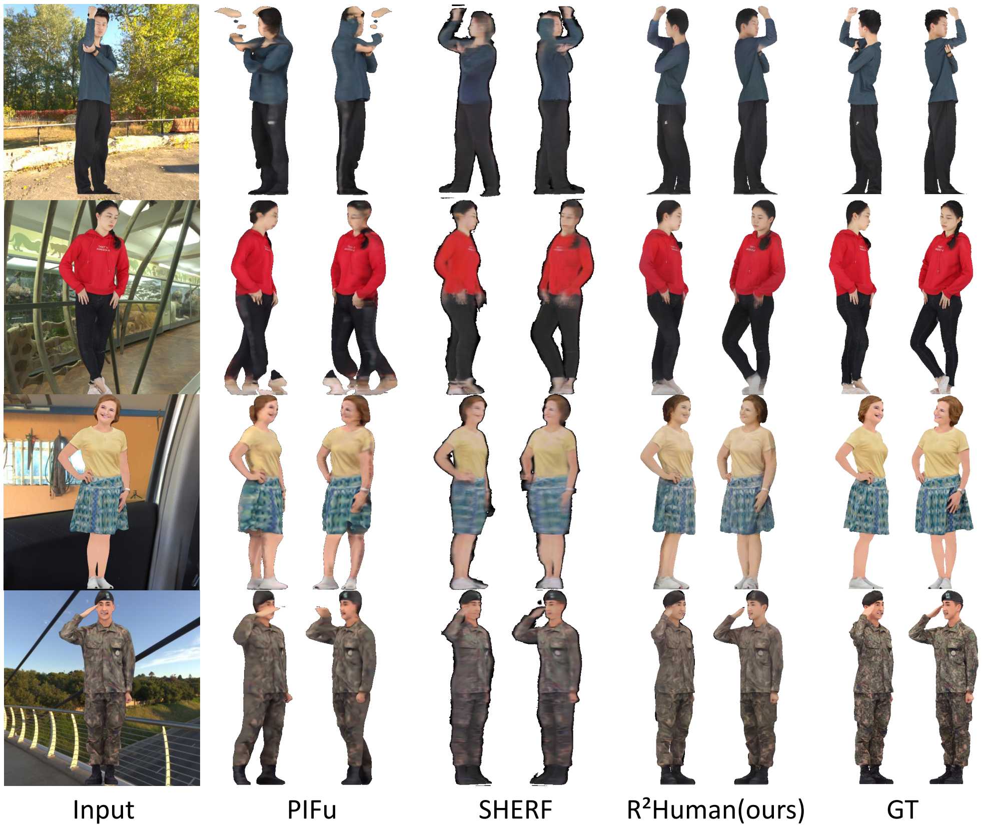

Figure 3: Novel view rendering on THuman2.0 (top tow rows) and 2k2k dataset (bottom tow rows).

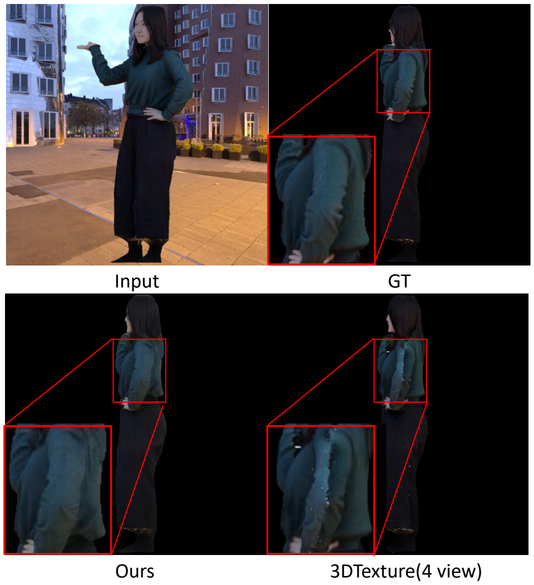

Figure 4: Qualitative comparison with 3DTexture. The results show that even with four views, our single-view approach still outperforms traditional texture mapping.

4.2 Ablation study

We conducted ablation experiments to compare variants of our architecture, different training strategies and different testing strategies.

Table 4: Comparison of Running Times.

	PIFu	SHERF	Ours	Ours (TensorRT)
Time (ms)	5741.62	1257.90	92.09	**13.13**

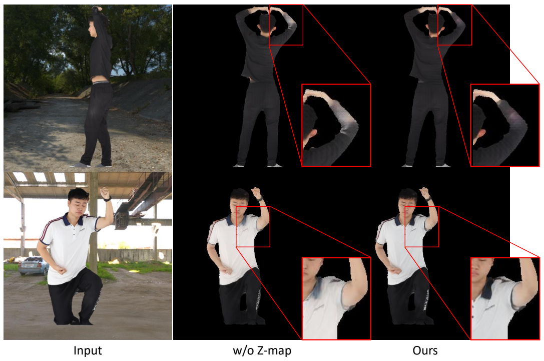

Figure 5: Ablation study comparing our model with and without the Z-map in the decoder.

Effects of Fourier occupation field (FOF) in the encoder. We remove the FOF input during encoding to verify the effects of 3D information on the results. Quantitative comparisons are in the sixth row of Tab. 5, showing that all three metrics of rendering results without FOF drop. As shown in Fig. 5, FOF can help image encoder better perceive 3D information, thereby reducing rendering artifacts and blurriness.

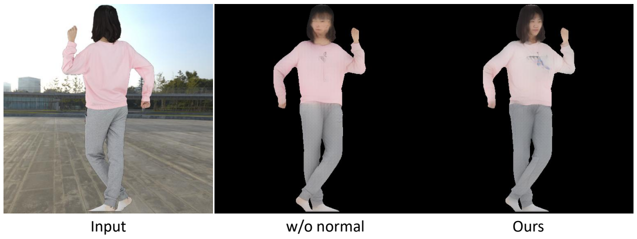

Figure 6: Ablation study comparing our model with and without the normal in the rendering network.

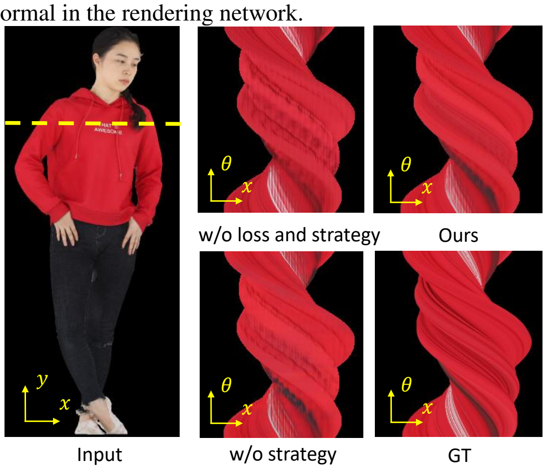

Figure 7: Ablation study on consistency loss in training and spatial fusion strategy in inference.

Effects of Z-map in the decoder. We retrained a variant without Z-map to explore the impact of Z-map. The first line of Tab. 5 shows the quantitative comparison results, and Fig. 5 shows the qualitative comparison results. It can be seen that Z-map can effectively improve the details of the generated results.

Effects of normal map in the rendering network. Fig. 6 illustrates the effects of incorporating normal maps into the rendering network. Normal maps are not generated by neural networks but are rendered from the reconstructed mesh. The second line of Tab. 5 shows the quantitative results. Normal maps can provide the network with more surface information, which allows the network to better obtain the correct color from the texture field and reduce the generation of artifacts.

Effects of FOF in the encoder. Fig. 8 shows the effect of using FOF in the encoder on the results, and the third line of Tab. 5 shows the quantitative results. FOF can help the encoder better understand the spatial information of the human mesh, thus enabling the network to synthesize accurate colors for large invisible areas, when the viewing angle changes substantially.

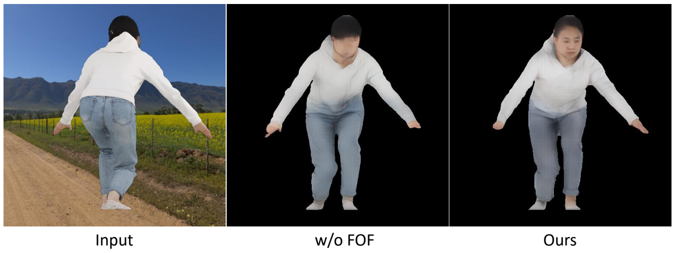

Figure 8: Ablation study comparing our model with and without the FOF in the encoder.

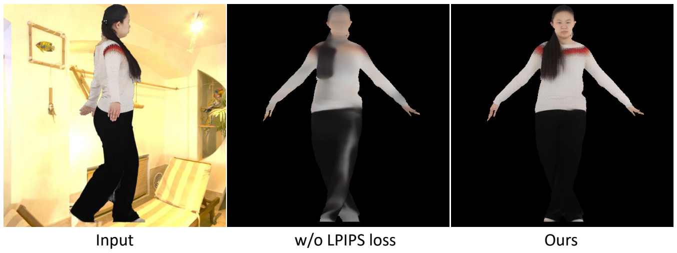

Figure 9: Ablation study comparing our model with and without the LPIPS loss.

Table 5: Quantitative evaluation for ablation study on THuman2.0 dataset.

Method	SSIM↑	PSNR↑	LPIPS↓
w/o Z_{map}	0.9570	30.18	0.0392
w/o normal	**0.9571**	30.26	0.0390
w/o FOF	0.9567	29.93	0.0394
w/o LPIPS loss	0.9382	25.81	0.0808
Ours(full)	**0.9571**	**30.52**	**0.0387**

Effects of LPIPS loss. To explore the effects of perceived loss on training, we retrained a variant without using the perceived loss. Fig. 9 compares results obtained with and without the use of perceived loss, and we also make a quantitative comparison in the fourth line of Tab. 5. The results show that the LPIPS loss greatly improves the network's ability to learn the texture information, resulting in photo-realistic rendering results.

Effects of consistency strategy. We add a consistency loss and a spatial fusion strategy to ensure the multi-view consistency of the results. To explore the impact of the spatial fusion strategy on the results, we retrain a variant without the consistency loss and use the method described in [28] to draw EPI images to qualitatively evaluate the view consistency. Fig. 7 compares the results with and without the spatial fusion strategy, and it can be seen that our spatial fusion strategy effectively enhances the view consistency during rotation.

4.3 Real-time Rendering System

Our pipeline, designed for real-time monocular human novel view synthesis, operates as follows: Images are captured from the video stream using OpenCV [4] and preprocessed to a resolution of 512×512. We then apply FOF-SMPL [5] to infer the occupation field. Subsequently, we employ the marching cube algorithm to extract the mesh and render the depth and normal maps. The final high-fidelity novel view images are rendered in real-time with the camera parameters Π^t and the previously processed data. Our pipeline, implemented with TensorRT on a single RTX-3090 GPU, achieves a performance of 24+ FPS, with the potential for further enhancement via additional GPUs. Fig. 10 showcase some of the real-time reconstruction results. Although our method trains to process each frame separately using synthetic data, it can still provide reasonable temporal consistency. Please refer to the demo video for more results.

4.4 Virtual Reality and Augmented Reality Applications

Our method has many meaningful applications, and we present two applications here. As shown in the top row of Fig. 11, we render the human appearance into a panoramic photo to generate the person in a virtual environment and display it in VR glasses. Specifically,

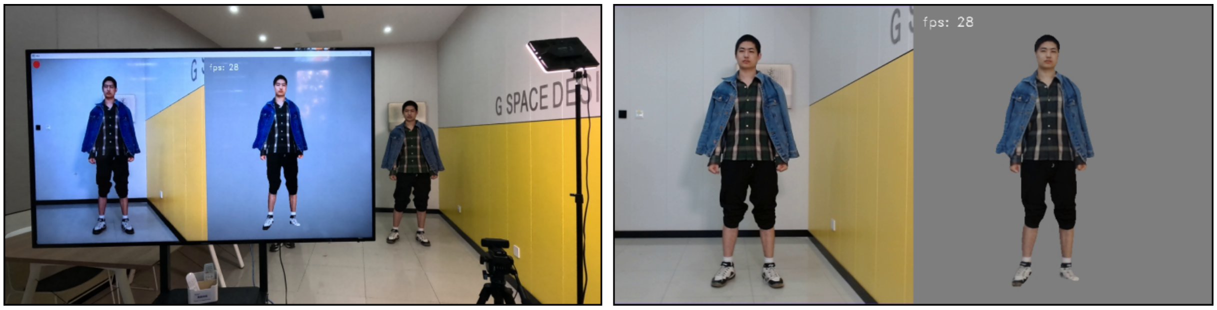

Figure 10: Real-time rendering results by our method.

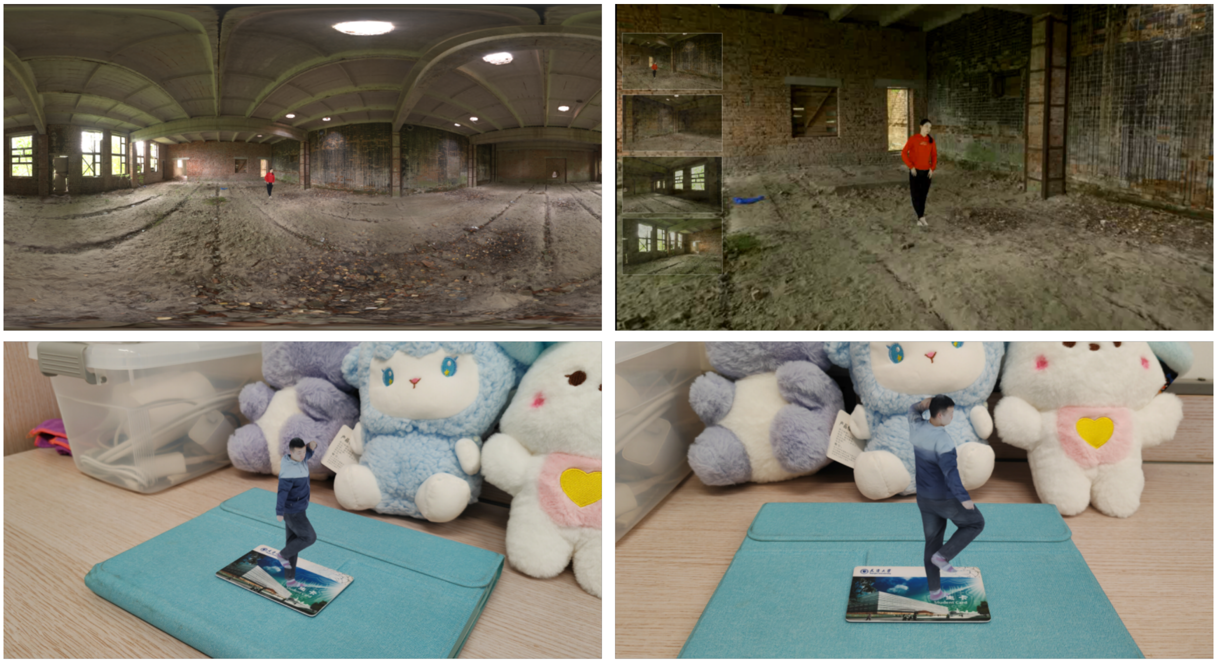

Figure 11: VR/AR rendering results by our method.

we project the results into a spherical coordinate system and expand the human appearance synthesized into the panoramic image to obtain the panoramic human appearance. By rendering the 3D human appearance into the virtual environment, our method provides people with a more realistic and stunning virtual experience. In addition, our method can also be integrated into the real scene of AR applications. As shown in the bottom row of Fig. 11, given the camera parameters of each frame in the video, we can render the human appearance in a specific scene and observe the 3D human from different views. Our VR/AR examples can generate many realistic applications in certain situations, *e.g.*, AR/VR education, immersion effects in games, virtual character communication in real scenes, etc. Also, our method of realistic 3D human appearance rendering has promising applications in immersive telepresence. In virtual conferences, the real 3D human appearance can greatly increase people's immersion. Applying our method in these VR/AR applications is a valuable direction in the future.

5 Conclusion and Discussion

Conclusion. Rendering 3D human appearance from a single image in real-time is important for achieving holographic communication and immersive VR/AR. We propose a novel method, combining the strengths of implicit texture field and explicit neural rendering, for real-time inference and rendering of realistic 3D human appearance from a monocular RGB image. We propose a new representation, Z-map, to alleviate depth ambiguities in rendering and enable high-fidelity color reconstruction. We also design a consistency loss and a spatial fusion strategy to ensure the multi-view coherence and reduce the jittering phenomenon. Experimental results show that the proposed method achieves state-of-the-art performance on both synthetic data and real-world images.

Broader Impact. Our proposed framework has the potential to significantly advance the development of holographic communication. However, high-fidelity novel view synthesis may also raise privacy concerns. Therefore, we strongly recommend that regulators establish ethical guidelines and regulatory frameworks that strike a balance between innovation and privacy protection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62122058 and 62171317), and the Science Fund for Distinguished Young Scholars of Tianjin (No. 22JCJQJC00040).

References

[1] B. AlBahar, S. Saito, H.-Y. Tseng, C. Kim, J. Kopf, and J.-B. Huang. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023. 1, 2

[2] T. Alldieck, M. Zanfir, and C. Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1506–1515, 2022. 1, 3

[3] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1034–1040. IEEE, 1997. 1

[4] G. Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 7

[5] Q. Feng, Y. Liu, Y.-K. Lai, J. Yang, and K. Li. Fof: learning fourier occupancy field for monocular real-time human reconstruction. *Advances in Neural Information Processing Systems*, 35:7397–7409, 2022. 2, 3, 5, 7

[6] Q. Feng, Y. Liu, Y.-K. Lai, J. Yang, and K. Li. Monocular real-time human geometry reconstruction. In *CAAI International Conference on Artificial Intelligence*, pp. 594–598, 2022. 2

[7] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12858–12868, 2023. 2

[8] S.-H. Han, M.-G. Park, J. H. Yoon, J.-M. Kang, Y.-J. Park, and H.-G. Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12869–12879, 2023. 5

[9] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9352–9364, 2023. 1, 2, 3, 5

[10] B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5605–5615, 2022. 2

[11] T. Jiang, X. Chen, J. Song, and O. Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16922–16932, 2023. 2

[12] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pp. 402–418. Springer, 2022. 2

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[14] J. Lawrence, D. B. Goldman, S. Achar, G. M. Blascovich, J. G. Desloge, T. Fortes, E. M. Gomez, S. Häberling, H. Hoppe, A. Huibers, et al. Project starline: A high-fidelity telepresence system. 2021. 2

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *Acm Transactions on Graphics*, 34(Article 248), 2015. 2, 3

[16] Y. Lu, H. Yu, W. Ni, and L. Song. 3d real-time human reconstruction with a single rgbd camera. *Applied Intelligence*, 53(8):8735–8745, 2023. 2

[17] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[18] P. Nguyen-Ha, N. Sarafianos, C. Lassner, J. Heikkilä, and T. Tung. Free-viewpoint rgb-d human performance capture and rendering. In *European Conference on Computer Vision*, pp. 473–491. Springer, 2022. 2

[19] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9054–9063, 2021. 1, 3

[20] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2304–2314, 2019. 1, 2, 5

[21] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020. 2

[22] U. Sara, M. Akter, and M. S. Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 5

[23] R. Shao, L. Chen, Z. Zheng, H. Zhang, Y. Zhang, H. Huang, Y. Guo, and Y. Liu. Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–10, 2022. 1, 2

[24] R. Shao, H. Zhang, H. Zhang, M. Chen, Y.-P. Cao, T. Yu, and Y. Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15872–15882, 2022. 1, 3

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[26] H. Song, B. Yoon, W. Cho, and W. Woo. Rc-smpl: Real-time cumulative smpl-based avatar body generation. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 89–98. IEEE, 2023. 2

[27] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 20–36, 2018. 2

[28] A. Vianello, J. Ackermann, M. Diebold, and B. Jähne. Robust hough transform based 3d reconstruction from circular light fields. In *CVPR*, 2018. 7

[29] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 836–850. Springer, 2014. 2, 5

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[31] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pp. 16210–16220, 2022. 2

[32] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black. Econ: Explicit clothed humans obtained from normals. *arXiv preprint arXiv:2212.07422*, 2022. 2

[33] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13286–13296. IEEE, 2022. 2

[34] X. Yang, Y. Luo, Y. Xiu, W. Wang, H. Xu, and Z. Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9122–9132, 2023. 2

[35] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5746–5756, 2021. 5

[36] M. Zhang, Q. Feng, Z. Su, C. Wen, Z. Xue, and K. Li. Joint2human: High-quality 3d human generation via compact spherical embedding of 3d joints. *arXiv preprint arXiv:2312.08591*, 2023. 2

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. 5

[38] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human recon-

struction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 2

[39] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7739–7749, 2019. 2

[40] T. Zhou, J. Huang, T. Yu, R. Shao, and K. Li. Hdhuman: High-quality human novel-view rendering from sparse views. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2

[41] W. Zielonka, T. Bagautdinov, S. Saito, M. Zollhöfer, J. Thies, and J. Romero. Drivable 3d gaussian avatars. 2024. 2