DoubleField: Bridging the Neural Surface and Radiance Fields for High-fidelity Human Reconstruction and Rendering

Ruizhi Shao[1], Hongwen Zhang[1], He Zhang[2], Mingjia Chen[1], Yan-Pei Cao[3], Tao Yu[1], Yebin Liu[1]
[1]Tsinghua University [2]Beihang University [3]Kuaishou Technology

Abstract

We introduce DoubleField, a novel framework combining the merits of both surface field and radiance field for high-fidelity human reconstruction and rendering. Within DoubleField, the surface field and radiance field are associated together by a shared feature embedding and a surface-guided sampling strategy. Moreover, a view-to-view transformer is introduced to fuse multi-view features and learn view-dependent features directly from high-resolution inputs. With the modeling power of DoubleField and the view-to-view transformer, our method significantly improves the reconstruction quality of both geometry and appearance, while supporting direct inference, scene-specific high-resolution finetuning, and fast rendering. The efficacy of DoubleField is validated by the quantitative evaluations on several datasets and the qualitative results in a real-world sparse multi-view system, showing its superior capability for high-quality human model reconstruction and photo-realistic free-viewpoint human rendering. Data and source code will be made public for the research purpose.

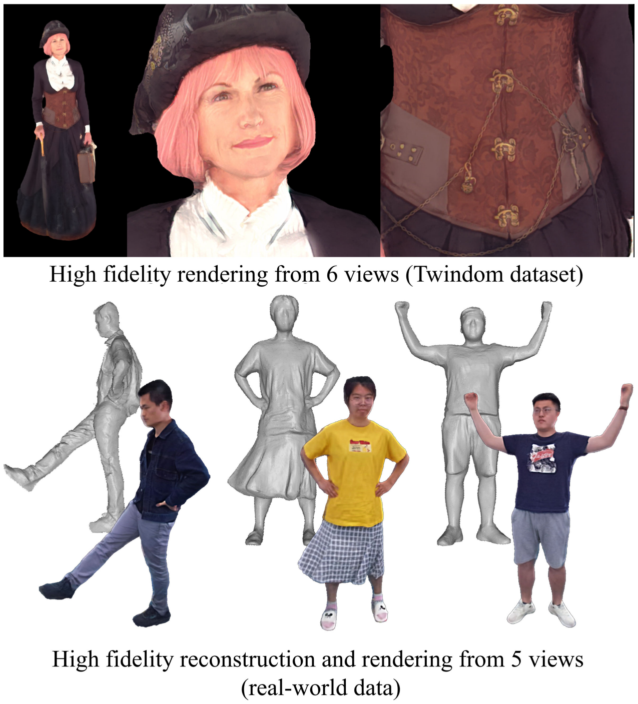
High fidelity rendering from 6 views (Twindom dataset)
High fidelity reconstruction and rendering from 5 views
(real-world data)

Figure 1. Given sparse multi-view RGB images, our method achieves high-fidelity human reconstruction and rendering.

1. Introduction

The surface fields [2, 31, 35] and the radiance fields [32, 63] have recently emerged as promising solutions for geometry modeling [12, 39, 40, 66] and texture rendering [37, 59, 65] of 3D human in an implicit and continuous manner, respectively. However, their limitations become apparent when considering simultaneous geometry and appearance reconstruction, not to say under sparse multi-view settings. Specifically, the surface fields [12, 39, 64, 67] separate the geometry learning from appearance learning and thus block the joint finetuning ability for more detailed geometry and rendering results. Moreover, the radiance fields [21, 32, 36, 37, 44] entangle the learning of geometry and appearance in an implicit manner without effective mutual constraints, leading to inconsistent geometry reconstruction and relatively low training efficiency. Despite the representations, the feature fusion strategy also dominates the final reconstruction quality when deploying the algorithms under multi-view setups, especially in the real-world systems. Even with high-resolution images as input, the limited representation power of features (feature map or feature volume) [37, 40] as well as the calibration and the geometry inference errors (especially for real captured data) will significantly deteriorate the detail reconstruction performance due to multi-view inconsistency for current implicit field based methods [37, 39, 64].

To overcome the limitations above for achieving high-quality 3D human reconstruction from sparse-view setups, we propose a novel DoubleField framework (to effectively bridge the surface and radiance fields and enable a shared learning space for both geometry and radiance reconstruction) and a view-to-view transformer (to build self attention between multi-view inputs and cross attention between the input views and the query viewpoints for multi-view feature fusion). Specifically, for DoubleField, we build associations between the surface and radiance fields by using a feature embedding shared by these fields in the network architecture and a surface-guided sampling strategy. Such a shared learning space allows the surface and radiance fields be benefited from each other. On the one hand, the surface field imposes a geometry constraint to the radiance field and encourages a more consistent density distribution for neural rendering. On the other hand, the radiance field enables more geometry details in the surface field via differentiable rendering. Moreover, the surface-guided sampling strategy disentangles the geometry component from the appearance modeling, so that DoubleField has a faster learning process while improving the reconstruction and rendering performances.

When deploying DoubleField with multi-view inputs, we propose a view-to-view transformer to build a self attention between multi-view inputs, and more importantly a cross attention between the input views and the query viewpoints. We achieve this by adopting an encoder-decoder architecture in our view-to-view transformer. Specifically, the encoder aims to fuse multi-view features while the decoder aims to produce view-dependent features based on the learned attention between the query view and all input views. Thanks to the attention learning ability of the transformer, the multi-view inconsistency issue is alleviated in our method, as the attention in the transformer handles the relationships between the input and the query views and is more robust to the geometry inference and calibration errors in real-world multi-view setups. Besides, the view-to-view transformer also enables our method to utilize the original high-resolution images. By taking the raw RGB values into accounts, the view-to-view transformer can directly learn the view-dependent features from high-resolution images and contribute to high-fidelity rendering performances.

In comparison with existing approaches [37, 39, 64] built upon surface and radiance fields, DoubleField not only improves the reconstruction quality of both geometry and appearance but also has the capability to eliminate the prerequisite SMPL fitting in previous methods [37] and even handle loose clothing (e.g., long dress) . More importantly, benefiting from the ability to leverage large dataset, DoubleField can fully utilize the priors in the large scale human scan dataset and achieve direct inference and fast finetuning for high-resolution free viewpoint rendering. In summary, Our contributions in this work are: 1) a DoubleField framework (a shared double embedding and a surface-guided sampling strategy) to combine the merits of both surface and radiance fields for sparse multiview human reconstruction and rendering; 2) a view-to-view transformer to fully utilize ultra-high-resolution image inputs in an efficient manner; 3) our method achieve state-of-the-art performance on both geometry reconstruction and texture rendering of human performances using sparse-view inputs.

2. Related Work

Neural implicit field Recently, neural implicit fields have emerged as powerful representations for geometry reconstruction and graphics rendering. Compared with the traditional explicit representations, such as meshes, volumes, and point clouds, neural implicit fields encode 3D models via neural networks that directly map 3D locations or viewpoints to the corresponding properties of occupancy [2, 31], SDF [35], volumes [27], and radiance [32] *etc*. Conditioned on spatial coordinates rather than discrete voxels or vertices, neural implicit field is continuous, resolution-independent, and more flexible, which enables higher quality surface recovery and photo-realistic rendering. For geometry reconstruction, methods based on surface fields [39, 40, 54] can generate detailed models from one or few images, and the high-fidelity geometry is achieved using local implicit field [1, 16]. For graphics rendering, methods based on implicit field are suitable for differentiable rendering [17, 24, 32, 42, 58]. Among them, the recently proposed NeRF [32] has made significant progress in novel view synthesis and photo-realistic rendering, which inspires many derivative methods [23, 30, 38, 41, 50, 59] and applications. Recently, there are also concurrent works [34, 49, 57] combine surface field and radiance field in an explicit manner and demonstrate promising results for case-specific learning and inference. However, extending them to large scale human scan dataset training for general inference is not straightforward. Explicitly building a clear numerical relationship between two fields is also limited to represent solid, non-transparent surfaces. In contrast, our DoubleField framework combines these two fields at the feature level in an implicit manner so that we can naturally incorporate pixel-aligned features and learn geometry prior from the large scale dataset during training. Our implicit combination is also more suitable to handle general and complicated human cases such as hair, semitransparent skirts, and thin clothes *etc*.

Human reconstruction Lately, there are numerous efforts devoted to capturing template-based human body from monocular or multi-view cameras at different levels, including shape and pose [14, 19, 22, 46, 62], and cloth surface [4, 7, 8, 48, 55]. Limited by the representation ability, these methods typically have low-quality results for both geometry and appearance recovery. Moreover, it is also difficult for those template-based algorithms to handle topology changes. Other approaches to high-quality human reconstruction require extremely expensive requirements such as dense viewpoints [18, 25, 26, 52] or even controlled lighting [3, 10]. Recently, implicit fields [15, 39, 64] enable detailed geometry reconstruction from sparse views. Based on

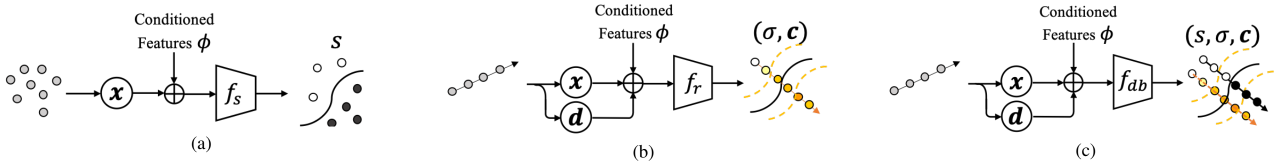

Figure 2. Comparison of different neural field representations. **(a)** Neural surface field in PIFu [39]. **(b)** Neural radiance field in PixelNeRF [59]. **(c)** The proposed DoubleField. The joint implicit fucntion f_{db} bridges the surface field and the radiance field.

sparse RGB-D cameras, the high-fidelity geometry reconstruction can be also achieved in real-time [11, 43, 53, 60]. Very recently, Peng *et al.* [37] propose to learn a neural radiance field with the guidance of a predefined template (*i.e.*, SMPL [28]) and achieve promising results on novel view synthesis from dynamic sequences. However, their method assumes the availability of an accurate estimation of the body template. Moreover, the simultaneous reconstruction of high-fidelity geometry and appearance from sparse-view input remains very challenging for existing solutions. Our work exploits a new path for high-quality geometry reconstruction and high-fidelity human rendering without the need of body templates.

Transformer Vaswani et al. [47] proposed Transformer, the first sequence transduction model based entirely on attention. The efficacy of Transformer is recently shown in a wide range of NLP and CV problems [5, 6, 61]. The attention mechanism, which is the core of transformer, has been proven by numerous literature to capture long-range dependencies [47, 51]. Its ability to obtain correlation has applied to many applications such as visual question answering [20], texture transferring [56], multi-view stereo [29], hand pose estimation [13], and human recontruction [64]. In our work, we apply a view-to-view transformer to capture the correspondences across the multi-view inputs.

3. Preliminary

Our DoubleField couples the representation power of the neural surface field [39] and the radiance field [32, 59]. In this section, we give a brief introduction of these two fields.

Neural Surface Field The neural surface field represented as the occupancy field [31, 39] is a resolution-independent representation for modeling 3D surface. As shown in Fig. 2a, a surface field can be formulated as an implicit function f_s mapping 3D points $\boldsymbol{x}$ to the surface field value s, e.g. $f_s(\boldsymbol{x}) = s : s \in [0, 1]$. To improve generalization and obtain detailed geometry, PIFu [39] conditions it on pixel-aligned image features using the following formulation:

$$f_s(\boldsymbol{x}, \phi(\boldsymbol{x}, \boldsymbol{I})) = s, \tag{1}$$

where $\phi(\boldsymbol{x}, \boldsymbol{I})$ is the image features located at the projection of $\boldsymbol{x}$ on the image $\boldsymbol{I}$. PIFu further extends this formulation to reconstruct texture on the surface by predicting RGB color $\boldsymbol{c}$ on the points $\boldsymbol{x}_c$ satisfied $f_s(\boldsymbol{x}_c) = 0.5$: $f_c(\boldsymbol{x}_c, \phi(\boldsymbol{x}_c, \boldsymbol{I})) = \boldsymbol{c}$. Though PIFu provides a straightforward solution for jointly modeling geometry and appearance, it isolates geometry and texture and makes the texture learning space discontinuous, hindering the geometry optimization process under texture supervisions [33].

Neural Radiance Field As shown in Fig. 2b, NeRF [32] represents a scene as a continuous volumetric radiance field f_r of the density σ and color $\boldsymbol{c}$, which describes geometry and appearance in an entangled form: e.g. $f_r(\boldsymbol{x}, \boldsymbol{d}) = (\sigma, \boldsymbol{c})$, where $\boldsymbol{d}$ is the viewing direction. Under this formulation, volumetric rendering can be used to synthesize novel view images by integrating along the projection rays:

$$\hat{C}(\boldsymbol{r}(t)) = \int_{t_n}^{t_f} T(t)\sigma(t)\boldsymbol{c}(t)dt, \tag{2}$$

where $\boldsymbol{r}(t) = \mathbf{o} + t\mathbf{d}$ denotes a camera ray with the origin $\mathbf{o}$ and direction $\mathbf{d}$. $T(t) = \exp\left(-\int_{t_n}^{t}\sigma(s)ds\right)$ tackles with occlusion, and $[t_n, t_f]$ is the pre-defined depth bounds. To achieve novel view synthesis from only sparse multi-view inputs, PixelNeRF [59] extends NeRF to leverage pixel-aligned image features in a similar manner to PIFu:

$$f_r(\boldsymbol{x}, \boldsymbol{d}, \phi(\boldsymbol{x}, \boldsymbol{I})) = (\sigma, \boldsymbol{c}). \tag{3}$$

Since the entangled modeling of density and color brings high flexibility for the training of NeRF, the surface learned in PixelNeRF is inconsistent given only sparse-view inputs, which leads to artifacts such as ghost-like or blurry results in novel view rendering. In addition, the highly flexible nature of the vanilla NeRF makes the training, and finetuning of its derivative solutions [37, 59] time-consuming.

4. Method

Our method is built on top of the DoubleField network and a view-to-view transformer. As illustrated in Fig. 3, given only sparse-view segmented images with ultra-high resolutions (e.g., 4K), our method can achieve both high-fidelity geometry and appearance reconstruction results without using any human body template.

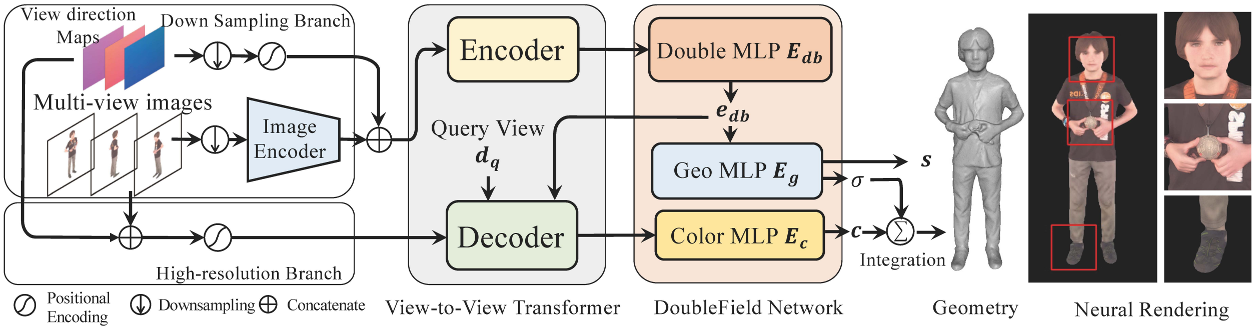

Figure 3. Pipeline of our method. Given sparse multi-view segmented images and the view direction maps, the encoder of our view-to-view transformer fuses low-resolution image features from different viewpoints and output the fused features. The double MLP E_{db} takes the fused features as inputs and produces the double embedding $\boldsymbol{e}_{db}$, which will be used to predict the surface field s and the density value σ by the geometry MLP. For the prediction of high-fidelity texture, the decoder takes the double embedding $\boldsymbol{e}_{db}$, query view direction $\boldsymbol{d}$, known views direction $\boldsymbol{d}^i$ and the colored encoding $\boldsymbol{p}(\boldsymbol{x})$ of the ultra-high-resolution images as inputs and produces the texture embedding $\boldsymbol{e}_c$ for the prediction of color values $\boldsymbol{c}$.

In this Section, we first introduce our DoubleField network by bridging the surface field and the radiance field in an implicit manner (Sect. 4.1). Based on DoubleField, an efficient view-to-view transformer is designed to leverage high-resolution images and adaptively synthesis photo-realistic rendering results (Sect. 4.2). Our network also supports efficient finetuning to recover high-fidelity geometry and appearance from high-resolution images (Sect. 4.3).

4.1. DoubleField Network

To overcome the limitations of existing neural field representations, we introduce the DoubleField network. The core of DoubleField consists of a shared embedding and a surface-guided sampling strategy, which connects the surface field and the radiance field so that they can be benefited from each other.

Basically, DoubleField can be formulated as a joint implicit function f_{db} represented by multi-layer perceptrons (MLPs) to fit both the surface field and the radiance field: $f_{db}(\boldsymbol{x}, \boldsymbol{d}) = (s, \sigma, \boldsymbol{c})$. Besides, DoubleField is also conditioned on pixel-aligned images features $\phi(\boldsymbol{x}, \boldsymbol{I})$. Specifically, as shown in Fig. 2c, given the query point $\boldsymbol{x}$, viewing direction $\boldsymbol{d}$ and images features $\phi(\boldsymbol{x}, \boldsymbol{I})$, our DoubleField network f_{db} learns a shared double embedding and predicts the surface field s, the density field σ and the texture field c simultaneously. Our DoubleField network is composed of a shared MLP (the *Double MLP* E_{db}) for double embedding e_{db} and two individual MLPs (the *geometry MLP* E_g and the *texture MLP* E_c) for the surface field and the radiance field prediction, as illustrated in Fig. 3. Overall, our DoubleField network can be formulated as:

$$\begin{aligned}\mathbf{e}_{db} &= E_{db}(\gamma(\boldsymbol{x}),\phi(\boldsymbol{x}, \boldsymbol{I})),\\ (s,\sigma) &= E_g(\mathbf{e}_{db}),\quad c = E_c(\mathbf{e}_{db}, \boldsymbol{d}),\\ f_{db}(\boldsymbol{x}, \boldsymbol{d}, \phi(\boldsymbol{x}, \boldsymbol{I})) &= (s, \sigma, \boldsymbol{c}),\end{aligned} \tag{4}$$

where $\gamma(\boldsymbol{x})$ is the positional encoding of $\boldsymbol{x}$, E_g is a *geometry MLP* for the prediction of occupancy s in the surface field and the density σ in the radiance field, while E_c is a *texture MLP* for prediction of the color c in the radiance field. Since s and σ are two output values of the last layer in the same MLP, such formulation implicitly builds a strong association between the two fields and enables their cooperation at the feature level.

Surface-guided Sampling Strategy To further facilitate the relation learned between the two fields and accelerate the rendering process, we make full use of the surface field and propose a surface-guided sampling strategy for DoubleField. The surface-guided sampling strategy will determine the intersection points in the surface field at first and then perform fine-grained sampling around the intersected surface. Specifically, given camera parameters of the rendering view and the ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$, a uniform sampling is firstly applied along the ray in the depth bounds $[t_n, t_f]$ with N_s sampling points, and each point is formulated as $\boldsymbol{x}_i = \mathbf{o} + t_i\mathbf{d}$. We query the surface field value of each point to determine the first intersection position $min\{t_i |\ s(\mathbf{o} + t_i\mathbf{d}) \geq 0.5\}$ on the surface. These intersections are then used to guide the sampling at a more fine-grained level by considering the radiance field surrounding the real surface in an interval of δ with N_r sampling points.

Our surface-guided sampling strategy can emphasize the relation between two fields around the mesh surface which facilitates the training and the finetuning process. Compared with NeRF sampling, our strategy is much fast on account of less sampling points needed for integration.

4.2. View-to-View Transformer

When applying DoubleField to multi-view inputs, we need to fuse the features from multi-view images. A straight forward solution is adopting a fusing strategy similar to PIFu [39] or PixelNeRF [59], where the pixel-aligned features will be extracted from the multi-view images and then fused together for DoubleField inference. Specifically, given image inputs $\{\boldsymbol{I}^i\}(i=1,2,...,n)$ from n viewpoints and the corresponding camera parameters, the image features are first extracted by the image encoder. For the query point $\boldsymbol{x}$, the pixel-aligned features $\phi^i(\boldsymbol{x},\boldsymbol{I}^i)$ on the image $\boldsymbol{I}^i$ are first obtained based on the projection of $\boldsymbol{x}$. These pixel-aligned features extracted from the multi-view images are then fused together as $\boldsymbol{\Phi}(\boldsymbol{x})$:

$$\begin{aligned} \Phi^i &= \oplus(\phi^i(\boldsymbol{x},\boldsymbol{I}^i),\boldsymbol{d}^i) \\ \boldsymbol{\Phi}(\boldsymbol{x}) &= \psi(\Phi^1,...,\Phi^n), \end{aligned} \tag{5}$$

where $\oplus(...)$ is a concatenation operator, $\phi^i(...)$ is the pixel-aligned features on the i-th viewpoint image, $\boldsymbol{d}^i$ is the viewing direction in the coordinate system of the i-th input viewpoint, and $\psi(...)$ is a feature fusion operation such as average pooling [39] or self-attention [64]. The fused features $\boldsymbol{\Phi}(\boldsymbol{x})$ can be taken as the conditioned features for DoubleField in Eq. 4 to predict the corresponding geometry and appearance in the query direction $\boldsymbol{d}_q$: $f_{db}(\boldsymbol{x},\boldsymbol{d}_q,\boldsymbol{\Phi}(\boldsymbol{x}))=(s,\sigma,\boldsymbol{c})$.

Although the above multi-view feature fusion methods can produce robust and plausible results, they only leverage the relatively low resolution image feature maps. Moreover, the geometry inference errors and the noises of the calibration in real-world data also significantly limit the quality of the final rendering results. To overcome this limitation, we propose a view-to-view transformer to directly take the raw RGB values from high-resolution images as input with both self attention and cross attention schemes.

Specifically, our view-to-view transformer adopts an encoder-decoder architecture that leverages the observations of the point $\boldsymbol{x}$ from all input views, and more importantly, the direction $\boldsymbol{d}_q$ of the query view to predict the color feature $\boldsymbol{e}_c$ for view-dependent rendering. In this way, our view-to-view transformer not only effectively fuses multi-view features in its **encoder** but also enables the cross attention between the query view and all the input views in its **decoder**, which differs from existing transformer-based fusion methods [64] that only use the transformer as an encoder for self-attention between input views. In the following, we present the encoder and decoder of our view-to-view transformer.

Encoder The goal of the encoder is to fuse the geometry features from multi-view inputs. It adopts the self-attention and feed-forward operation ψ in Eq. 5 to obtain the fused features $\boldsymbol{\Phi}$, which will be fed into the *double MLP* E_{db} for the generation of the double embedding:

$$\begin{aligned} Q^e, K^e, V^e &= F^e_{Q,K,V}(\phi^1,...,\phi^n) \\ \Phi &= F^e(Att(Q^e,K^e,V^e)) \\ \boldsymbol{e}_{db} &= E_{db}(\gamma(\boldsymbol{x}),\boldsymbol{\Phi}), \end{aligned} \tag{6}$$

where $F^e_{Q,K,V}$ denotes the linear layers producing the query, key and value matrices Q^e, K^e, V^e, respectively, F^e is the feed-forward layer, and Att is the multi-head attention operation in the transformer.

Decoder The goal of the decoder is to produce the view-dependent color embedding $\boldsymbol{e}_c$ according to the observations from all input views, and the query view direction $\boldsymbol{d}_q$. To leverage the high-resolution information, the decoder takes both low- and high-level observations into account, including the raw rgb $\boldsymbol{p}^i$ and double embedding $\boldsymbol{e}_{db}$. Specifically, the process can be formulated as:

$$\begin{aligned} Q^d &= F^d_Q(\boldsymbol{d}_q) \\ K^d &= F^d_K(\boldsymbol{d}^1,...,\boldsymbol{d}^n) \\ V^d &= F^d_V([\boldsymbol{e}_{db},\gamma(\boldsymbol{p}^1)],...,[\boldsymbol{e}_{db},\gamma(\boldsymbol{p}^n)]) \\ \boldsymbol{e}_c &= F^d(Att(Q^d,K^d,V^d)) \end{aligned} \tag{7}$$

where F^d_Q, F^d_K, F^d_V denote the linear layers producing the query, key and value matrices Q^d, K^d, V^d, respectively, F^d is the feed-forward layer. Here, similar to the position encoding $\gamma(\boldsymbol{x})$, we also map the raw RGB values $\boldsymbol{p}^i$ to a higher dimensional space as the colored encoding $\gamma(\boldsymbol{p}^i)$ for the learning of high-frequency appearance variations [45].

After obtaining the color embedding $\boldsymbol{e}_c$ from the decoder, the high-resolution color at the point $\boldsymbol{x}$ is predicted by the *texture MLP* E_c: $\boldsymbol{c}=E_c(\boldsymbol{e}_c)$.

4.3. Training and Finetunning

Though our network can leverage high-resolution images as input, the expensive training time cost on such a high-resolution domain is unacceptable. For a more feasible solution, in implementation we divide the problem into two phases: low-resolution large-scale-dataset pre-training and efficient person-specific high-resolution finetuning.

Large-Scale Dataset Pre-training Our pre-training phase is similar with the training process of PIFu [39] and PixelNeRF [59]. We collect human models from Twindom[1] dataset (1,500 for training) and render low-resolution images with the size of 512 × 512. We adopt the spatial sampling strategy in PIFu [39] for the learning of geometry, and the proposed surface-guided sampling strategy for the learning of appearance. For the loss of geometry training,

[1] https://web.twindom.com/

we adopt the spatial sampling loss function in PIFu [39] and the implicit geometric regularization loss (L1 form) [9]:

$$\begin{aligned} L_g &= \frac{1}{N_g}\sum_{i=1}^{N_g}\|s(\boldsymbol{x_i}) - s^*(\boldsymbol{x_i})\|_2^2 \\ L_r &= \frac{1}{N_r}\sum_{i=1}^{N_r}\|\nabla s(\boldsymbol{x_i}) - n^*(\boldsymbol{x_i})\|_1, \end{aligned} \tag{8}$$

where $s^*(\boldsymbol{x_i})$ is the ground truth occupancy of $\boldsymbol{x_i}$, and $n^*(x_i)$ is the ground truth normal of $\boldsymbol{x_i}$. N_g and N_r are the number of sampling points for spatial sampling and geometric regularization, respectively. The regularization loss can further improve the quality of geometry reconstruction without requirement of normal map as input. To obtain the ground truth of normal, we only sample points on the mesh surface when applying regularization loss. And for appearance loss, we adopt the L1 loss between the rendered color and the ground truth color as:

$$L_c = \frac{1}{N_c}\sum_{i=1}^{N_c}|\hat{C}(\boldsymbol{r}_i) - C^*(\boldsymbol{r}_i)|, \tag{9}$$

where the rendered color $\hat{C}(\boldsymbol{r}_i)$ is obtained using the integration [32] along the ray $\boldsymbol{r}_i$ in the interval around the surface. $C^*(\boldsymbol{r}_i)$ is the ground truth color of ray $\boldsymbol{r}_i$. N_c is the number of sampling rays. In summary, our final loss can be formulated as: $L = \lambda_g L_g + \lambda_r L_r + \lambda_c L_c$, where λs balance the loss terms.

Finetuning Phase In the finetuning phase, the network takes the ultra-high-resolution images from the sparse multi-view of a specific human as input and finetune the network parameters in a self-supervised manner using differentiable rendering loss. Specifically, We first fix the transformer and the color MLP to finetune geometry for 2000 iterations and then fix the double MLP and the geometry MLP to finetune the color MLP for another 2000 iterations. In each iteration, we randomly select one view as ground truth and regard the other views as input. The only one loss function we used is Eq. 9 and the learning rate is tune down for stable finetuning performance (1e-6 in finetuning and 1e-5 in pre-training).

5. Experiment

5.1. Experiments on Synthetic Data

We evaluate our method by using synthetic rendering of multiview images on two high-quality 3D human scan datasets: 1) Twindom dataset (200 for testing), 2) THuman2.0 [60], a publicly-available high-quality human model dataset (100 for testing).

We compare DoubleField with the state-of-the-art approaches built upon the surface field and the radiance field, including PIFu [39], PixelNeRF [59], NeuralBody [37], and PIFuHD [40]. We also implement DVR [33] based on PIFu (denoted as PIFu+DVR) to validate the efficiency of the DoubleField representation and its finetuning ability on unseen data. For fair comparisons, we additionally train PIFu with regularization loss (PIFu+R) and replace the average pooling operation in PIFu [39], PIFuHD [40] and PixelNeRF [59] with self-attention modules for multi-view feature fusion. We retrain these networks with the same train-

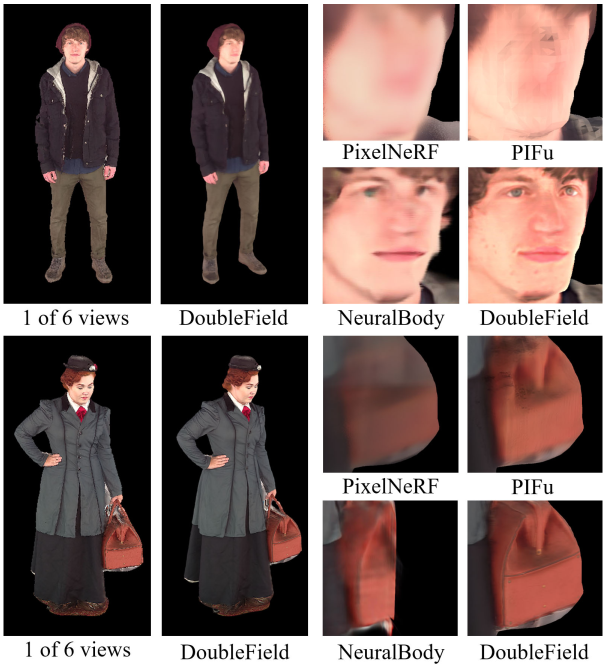

Figure 4. Comparison on appearance reconstruction using the Twindom dataset. PixelNeRF [59] and our method are finetuned with additional 4,000 iterations. Note that NeuralBody [37] can not handle additional objects which are far away from the human b

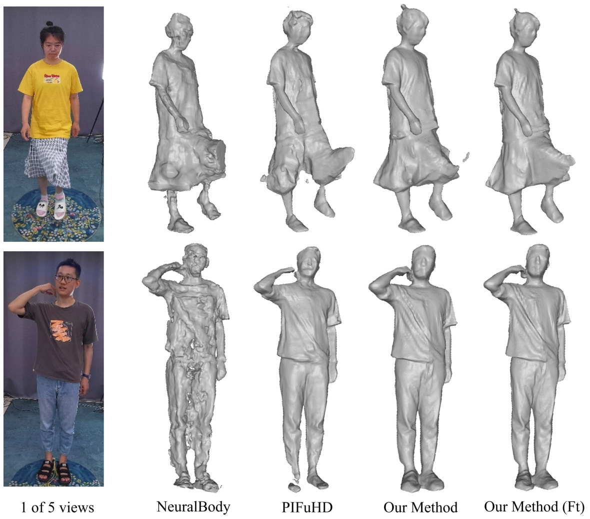

Figure 5. Comparisons of geometry reconstruction results using real multi-view image (5 views).

Method	Twindom (6 views Geo.) Chamfer	 P2S	THuman2.0 (6 views Geo.) Chamfer	 P2S
PIFu [39]	0.754	0.716	0.710	0.613
PIFu+R	0.739	**0.699**	0.697	0.606
PIFuHD [40]	0.742	0.701	0.700	0.609
PIFu+DVR [33]	0.746	0.701	0.709	0.611
PixelNeRF [59]	0.945	0.931	0.815	0.725
Our Method (w/o Ft)	**0.737**	0.700	**0.696**	**0.605**
NeuralBody [37]	1.597	2.146	1.528	2.126
PIFu+DVR (Ft)	0.779	0.736	0.724	0.623
PixelNeRF (Ft)	1.072	1.052	0.790	0.701
Our Method (Ft)	**0.711**	**0.690**	**0.662**	**0.589**

Table 1. Quantitative human geometry reconstruction results. Ft denotes the approaches finetuned with 4,000 iterations.

Method	Twindom (6 views Col.) PSNR	 SSIM	THuman2.0 (6 views Col.) PSNR	 SSIM
PIFu [39]	20.80	0.805	22.35	0.846
PIFu+DVR [33]	20.65	0.804	22.17	0.843
PixelNeRF [59]	21.57	0.808	22.95	0.854
Our Method (w/o Ft)	**22.95**	**0.842**	**24.23**	**0.880**
NeuralBody [37]	20.69	0.808	22.65	0.862
PIFu+DVR (Ft)	21.62	0.812	23.08	0.855
PixelNeRF (Ft)	21.85	0.813	23.57	0.863
Our Method (Ft)	**23.56**	**0.857**	**25.10**	**0.905**

Table 2. Quantitative human rendering results. Ft denotes the approaches finetuned with 4,000 iterations.

ing settings and datasets.

Comparisons on Geometry Reconstruction. For the comparison with NeuralBody [37], we regard NeuralBody as a frame-based method and train it on 6 viewpoint inputs for 15 hours. Due to the expensive training cost, we randomly pick only 50 models from Twindom test dataset and 30 models from THuman2.0 dataset for NeuralBody evaluation. We quantitatively evaluate the geometry recovery performance using the point-to-surface distance and the chamfer distance in Table. 2. Our method without finetuning achieves competitive results compared with PIFuHD, PIFu+R and PIFu+DVR. After finetuning, our method can further improve the quality of geometry even without ground truth geometry for supervision based on the DoubleField representation.

Comparisons on Appearance Rendering. To evaluate the appearance rendering performance, we prepare images of 4K resolution rendered from 30 viewpoints, and use images from 6 fixed viewpoints as input and images from other 24 views for evaluation. Quantitative results are shown in Table. 2. Benefiting from the view-to-view transformer and the DoubleField representation, our method achieves high-fidelity rendering. Moreover, our method can support higher quality appearance reconstruction with quick finetuning in 20 minutes (10 minutes for geometry finetuning and 10 minutes for texture and transformer finetuning, 4,000 iterations in total). Moreover, our method generalizes well to scenarios like loose clothes (e.g. long skirts) and object interactions as shown in Fig. 4.

	Twindom (6 views Col.) PSNR	 SSIM	THuman2.0 (6 views Col.) PSNR	 SSIM
average pooling	22.53	0.826	23.89	0.870
w/o DbMLP	22.01	0.818	23.42	0.866
w/o CE	22.89	0.831	24.11	0.874
Our method (w/o Ft)	22.95	0.842	24.23	0.880
Ft w/o HD pixel	23.28	0.847	24.97	0.896
Our method (Ft)	23.56	0.857	25.10	0.905

Table 3. Ablation study on Twindom and Thuman2.0 dataset with four settings: Average pooling (use the same multi-view feature fusion in PIFu and PixelNeRF), W/o DbMLP (remove Double MLP and learn two fields separately), W/o CE (removes the color encoding and directly adopt 3-dim RGB), Ft w/o HD pixel (finetune using only low-resolution images).

Ablation Study. We compare different factors that contribute to our method. As shown in Tab. 3, compared with the view-to-view transformer and the color encoding, the DoubleField network has the most significant contribution to the final results. Meanwhile, the view-to-view transformer is more effective for achieving multi-view and cross view feature fusion than a simple pooling layer. We also conduct experiments in the high-resolution domain with finetuning. The model of "Ft w/o HD pixel" is finetuned using only low-resolution images (512x512). The performance of such setting is worse than our method but better than the others, demonstrating the ability of our view-to-view transformer to capture correspondences across different views and leverage the high resolution input.

5.2. Results on Real World Multi-view Data

We evaluate our geometry reconstruction and texture rendering performance using real-world data captured from sparse multi-view cameras (5 views). Fig. 5 compares the qualitative geometry reconstruction results of NeuralBody [37], PIFuHD [40], and our method. Note that our method is finetuned with the multi-view images at one frame, while NeuralBody [37] is trained with the whole mutli-view video sequence as it fails in the geometry reconstruction when only one frame is given. As show in Fig. 5, unlike NeuralBody [37], the surface reconstructed by our method is more consistent and contains more details. The finetuning can further fix some missing parts on the geometry such as holes, which shows that the double MLP has learned to build an implicit association between the two fields. Finally, even without using the normal maps as input, our method produce more accurate results compared with the multi-view extension of PIFuHD.

We further evaluate the rendering quality on the ZJU-mocap dataset [37] and our multi-view system. The results are shown in Fig. 6 and Fig. 7. Our method pro-

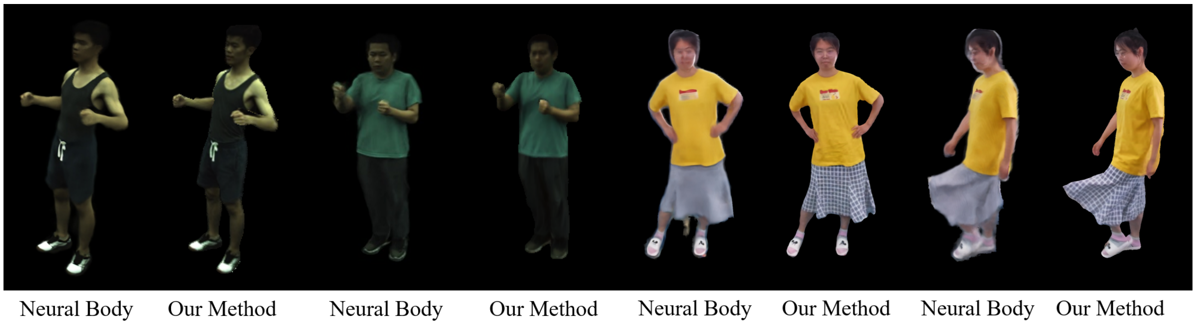

Figure 6. Comparisons with NeuralBody. 4 images on the left are from the ZJU-mocap dataset, and 4 images on the right are from a real world multi-view (5 views) system. Each video has 300 frames and we train NeuralBody for 20 hours.

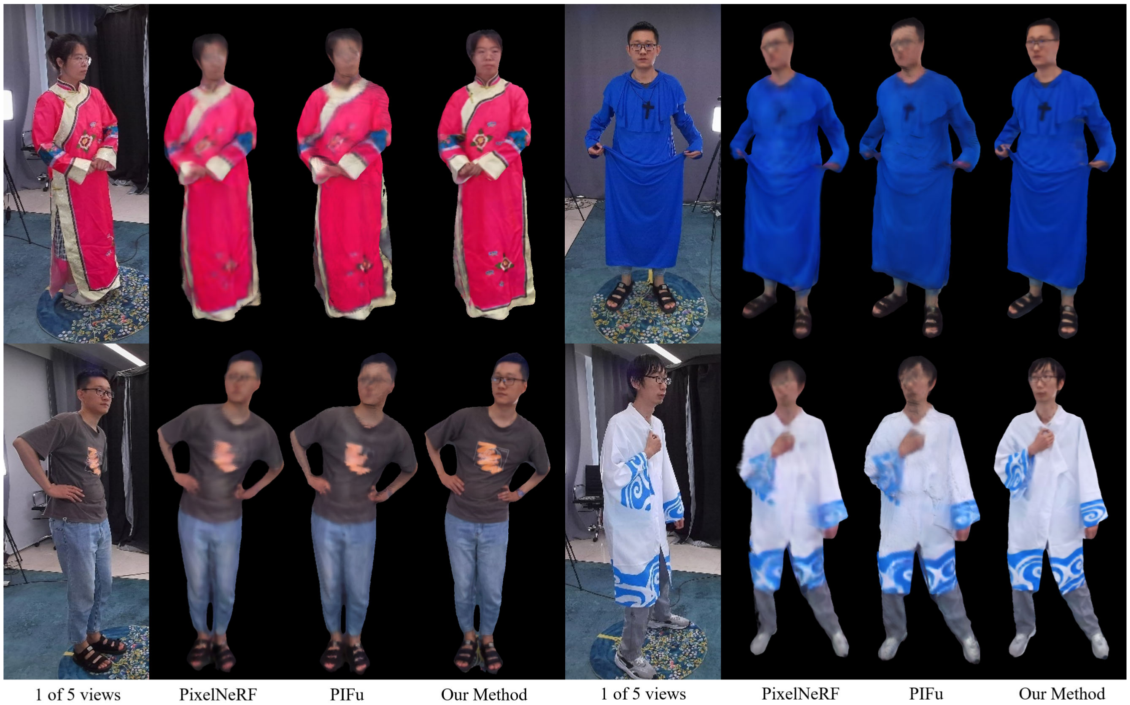

Figure 7. Comparisons on real world data under 5-views setting with PixelNeRF [59] and PIFu [39]

duces more clear rendering results using much less time for network finetuning (< 20 minutes V.S. > 15 hours). Moreover, our method does not rely on human shape prior SMPL [28] compared with NeuralBody [37] and achieves photo-realistic rendering even under challenging scenarios like swinging skirt, topological changes and loose cloth, which demonstrates the strong generalization capacity of our method to real world data. For more results, please refer to our supplementary video.

6. Discussion

Conclusion. We propose DoubleField to combine the merits of both geometry and appearance fields for human surface reconstruction and rendering under sparse view inputs. In our work, the proposed DoubleField network and view-to-view transformer enable a substantial performance improvement on both geometry reconstruction and texture rendering of human performances. We believe our approach can enlighten the follow-up works in the field of human rendering and reconstruction.

Limitations. The proposed pipeline still relies on accurate background image subtraction for Doublefield inference due to the requirement of pixel-aligned image feature extraction. Moreover, our method does not support reconstruction and rendering of multiple character scenarios.

Potential Social Impact. Our method focuses on free-viewpoint rendering of a human performance and can be used in sport games, movie, virtual reality, tele-presence, etc., which has no obvious negative societal impact.

Acknowledgement. This paper is sponsored by National Key R&D Program of China (2021ZD0113503) and the NSFC No. 62125107 and No. 62171255.

References

[1] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3D reconstruction. In *ECCV*, pages 608–625. Springer, 2020. 2

[2] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 1, 2

[3] 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. *TOG*, 34(4):1–13, 2015. 2

[4] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *TOG*, pages 1–10, 2008. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3

[7] Mingsong Dou, Henry Fuchs, and Jan-Michael Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *ISMAR*, pages 99–106. IEEE, 2013. 2

[8] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753. IEEE, 2009. 2

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, pages 3789–3799. PMLR, 2020. 6

[10] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *TOG*, 38(6):1–19, 2019. 2

[11] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM TOG*, 36(4):1, 2017. 3

[12] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *CVPR*, 2021. 1

[13] Lin Huang, Jianchao Tan, Jingjing Meng, Ji Liu, and Junsong Yuan. Hot-net: Non-autoregressive transformer for 3D hand-object pose estimation. In *ACM MM*, pages 3136–3145, 2020. 3

[14] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, pages 421–430. IEEE, 2017. 2

[15] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, pages 336–354, 2018. 2

[16] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3D scenes. In *CVPR*, pages 6001–6010, 2020. 2

[17] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3D shape optimization. In *CVPR*, pages 1251–1261, 2020. 2

[18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329. IEEE, 2018. 2

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, volume 31, pages 1564–1574, 2018. 3

[21] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *NeuriPS*, 2021. 1

[22] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *CVPR*, pages 4352–4362, 2019. 2

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *NeurIPS*, 2020. 2

[24] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, pages 2019–2028, 2020. 2

[25] Yebin Liu, Xun Cao, Qionghai Dai, and Wenli Xu. Continuous depth estimation for multi-view stereo. In *CVPR*, pages 2121–2128. IEEE, 2009. 2

[26] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE TVCG*, 16(3):407–418, 2009. 2

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

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 3, 8

[29] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *CVPR*, pages 1590–1599, 2020. 3

[30] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 2

[31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 1, 2, 3

[32] 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 *ECCV*, pages 405–421. Springer, 2020. 1, 2, 3, 6

[33] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *CVPR*, pages 3504–3515, 2020. 3, 6, 7

[34] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[35] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, June 2019. 1, 2

[36] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. In *ICCV*, 2021. 1

[37] 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 *CVPR*, 2021. 1, 2, 3, 6, 7, 8

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[39] 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 *ICCV*, October 2019. 1, 2, 3, 5, 6, 7, 8

[40] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR*, 2020. 1, 2, 6, 7

[41] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3D-aware image synthesis. In *NeurIPS*, volume 33, 2020. 2

[42] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[43] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, pages 246–264, 2020. 3

[44] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *CVPR*, pages 6226–6237, 2021. 1

[45] Matthew. Tancik, Pratul. P. Srinivasan, Ben. Mildenhall, Sara. Fridovich-Keil, Nithin Raghavan, Utkarsh. Singhal, Ravi. Ramamoorthi, Jonathan. T. Barron, and Ren. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. 2020. 5

[46] Yating Tian, Hongwen Zhang, Yebin Liu, and limin Wang. Recovering 3D human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923*, 2022. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[48] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *TOG*, pages 1–9, 2008. 2

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021. 2

[50] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3

[52] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *CVPR*, pages 1682–1691, 2020. 2

[53] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. *IEEE TPAMI*, 42(10):2508–2522, 2019. 3

[54] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3D reconstruction. In *NeurIPS*, 2019. 2

[55] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *TOG*, 37(2):1–15, 2018. 2

[56] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, pages 5791–5800, 2020. 3

[57] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34, 2021. 2

[58] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 2

[59] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7, 8

[60] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 3, 6

[61] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from

scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021. 3

[62] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 2

[63] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[64] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, 2021. 1, 2, 3, 5

[65] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022. 1

[66] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021. 1

[67] Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, and Tony Tung. Learning implicit 3D representations of dressed humans from sparse views. In *3DV*, 2021. 1