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

Ruizhi Shao[1], Hongwen Zhang[1], He Zhang[2], Yanpei Cao[3], Tao Yu[1], and Yebin Liu[1]

[1]Tsinghua University [2]Beihang University [3]Kuaishou Technology

High-fidelity rendering from 6 views High-fidelity rendering from 5 views

Figure 1: Given only sparse multi-view RGB images (6 views for the left on the Twindom dataset, 5 views for the right on the real-world data), our method achieves high-fidelity neural human rendering for loose clothes and various poses.

Abstract

We introduce DoubleField, a novel representation combining the merits of both surface field and radiance field for high-fidelity human rendering. Within DoubleField, the surface field and radiance field are associated together by a shared feature embedding and a surface-guided sampling strategy. In this way, DoubleField has a continuous but disentangled learning space for geometry and appearance modeling, which supports fast training, inference, and finetuning. To achieve high-fidelity free-viewpoint rendering, DoubleField is further augmented to leverage ultra-high-resolution inputs, where a view-to-view transformer and a transfer learning scheme are introduced for more efficient learning and finetuning from sparse-view inputs at original resolutions. 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 photo-realistic free-viewpoint human rendering. For code and demo video, please refer to our project page: `http://www.liuyebin.com/dbfield/dbfield.html`.

1 Introduction

The surface fields [28, 25, 2] and radiance fields [26, 48] have recently emerged as promising solutions to model 3D geometry and appearance in a resolution-independent and continuous manner. Though significant progress has been made towards detailed geometry recovery [32, 33, 50, 10] and photo-realistic texture recovery and rendering [45, 29], their limitations become apparent when considering the simultaneous reconstruction of both geometry and appearance.

Preprint. Under review.

	Training	Geometry	Rendering	Finetuning	Supervision	Inference
PIFu [32]	fast	high	traditional	not support	3D scan data	fast
PixelNeRF [45]	very slow	low	neural	slow	images only	slow
DoubleField	fast	high	neural	fast	images/3D	fast

Table 1: A brief comparison between our work and PIFu [32] and PixelNeRF [45].

The limitations of the existing surface fields and radiance fields originate from the trade-off between the continuity and disentanglement properties. Specifically, the reconstruction approaches [32, 27, 20] built upon surface fields typically learn texture on the surface, resulting in the distribution of the predicted texture highly concentrated on the surface. Such a narrow texture field is typically discontinuous and hinders the optimization processes of differentiable rendering. By contrast, the radiance fields [26, 48] enable the learning of a continuous texture field, but its geometry is entangled with the texture and lacks enough constraints. Such a geometry-appearance entanglement not only leads to the inconsistency and artifacts in the geometry recovery, especially under sparse multi-view settings, but also makes the training and inference of the radiance fields very time-consuming [26].

To overcome the limitations of existing neural field representations, we propose a novel DoubleField framework to bridge surface and radiance fields and enable a continuous but disentangled space for geometry and appearance learning. Specifically, we build associations between surface and radiance fields from the aspects of the network architecture and sampling strategy. 1) In our network architecture, an intermediate MLP learns a shared double embedding for both the two fields. The shared learning space facilitates the update of both fields with the back-propagated gradients so that both geometry and appearance can be learned in a continuous manner. 2) A surface-guided sampling strategy is proposed to first sample sparse points to determine the intersected surface and then sample dense points around the surface for volume rendering in the radiance field. Such a strategy imposes the geometry constraint for the radiance field and disentangles the geometry component from the appearance modeling, which not only accelerates the learning process but also improves the quality and consistency of the free viewpoint rendering results. With the proposed architecture and sampling strategy, DoubleField combines the merits of the two fields and naturally supports efficient finetuning on new data with self-supervision signals based on differentiable rendering.

To fully exploit the potential power of DoubleField, we take one more step forward to leverage ultra-high-resolution inputs. Instead of learning on coarse image features only, DoubleField is further augmented with a view-to-view transformer to directly take the raw RGB values of the images at the original resolution as inputs. This is motivated by the observation that the free-viewpoint rendering can be regarded as a view-to-view problem, i.e., generating novel-view images given sparse-view images, which is reminiscent of the text-to-text problem of a typical task in NLP [31]. For more efficient modeling of the high-fidelity appearance, we adopt a transductive transfer learning scheme for our network. Specifically, our network is first trained on the low-resolution pre-training task for the learning of general multi-view prior and then transferred and adopted to the high-fidelity domain by fast finetuning when dealing with ultra-high-resolution inputs. However, this is non-trivial since finetuning on sparse-view images is prone to overfitting. To overcome this issue, we conduct comprehensive experiments to measure the influence of different modules and empirically introduce a bottom-up finetuning strategy that can avoid overfitting with fast convergence. The experimental results on human reconstruction from sparse-view inputs demonstrate the state-of-the-art performance and high-fidelity rendering quality of our approach. The comparison of the proposed DoubleField representation with existing ones is summarized in Table 1. Our contributions in this work are listed as follows:

1. We propose a novel representation DoubleField to combine the merits of both surface and radiance fields. We bridge these two fields via a shared double embedding and a surface-guided sampling strategy so that DoubleField has a continuous but disentangled learning space for geometry and appearance modeling.

2. We further augment DoubleField to support high-fidelity rendering by introducing a view-to-view transformer to take the raw RGB values of the ultra-high-resolution images as inputs. The view-to-view transformer learns the texture mapping from the known viewpoints to the query viewpoints on the ultra-high-resolution domain.

3. We exploit a transfer learning scheme and a bottom-up finetuning strategy for more efficient training of our network so that it can have a fast convergence speed while avoiding the overfitting

issue. In this way, DoubleField can produce high-fidelity free-view rendering results given only sparse-view inputs, which demonstrates significant performance improvements upon prior work.

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, volumns, 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 [25, 2], SDF [28], volumes [21], and radiance [26] 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 [32, 33, 41] can generate detailed models from one or few images, and the high-fidelity geometry is achieved using local implicit field [14, 1]. For graphics rendering, methods based on implicit field are suitable for differentiable rendering [20, 44, 15, 35, 26]. Among them, the recently proposed NeRF [26] has made significant progress in novel view synthesis and photo-realistic rendering, which inspires many derivative methods [45, 24, 34, 38, 19, 30] and applications.

Multi-view human reconstruction There are numerous efforts devoted to capturing template-based human body from multi-view cameras at different levels, including shape and pose [12, 18], and cloth surface [4, 37, 8, 7, 42]. 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 dependencies such as dense viewpoints [16, 40] or even controlled lighting [3, 9]. Recently, implicit fields [13, 49, 32] enable detailed geometry reconstruction from sparse views. Based on sparse RGB-D cameras, the high-fidelity geometry reconstruction can be also achieved in real-time [46]. Very recently, Peng et al. [29] propose to learn a neural radiance field with the guidance of a predefined template (i.e., SMPL [22]) 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 inputs remains very challenging for existing solutions. Our work exploits a new path for high-fidelity human rendering without the need of body templates.

Transformer The efficacy of Transformer is recently shown in a wide range of NLP and CV problems [5, 6, 47]. The attention mechanism, which is the core of transformer, has been proved by numerous literature to capture long-range dependencies [36, 39]. Its ability to obtain correlation has applied to many applications such as visual question answering [17], texture transferring [43], multi-view stereo [23], and hand pose estimation [11]. Besides, transfer learning based on transformer [5] has made significant progress in NLP and showed great potential for generalization. In our work, we regard the free-viewpoint rendering problem as a view-to-view problem and apply a transformer to capture the correspondences across the multi-view inputs. Motivated by previous work, we adopt a transfer learning scheme to tackle the learning issue on ultra-high resolution images.

3 Preliminary

Our DoubleField representation is built upon the neural surface fields [32] and radiance field [26, 45]. In this Section, we briefly introduce the background of these two fields. Please aslo refer to Fig. 2 for the comparison of different neural field representations.

Neural Surface Field The neural surface field represented as the occupancy field [25, 32] is a resolution-independent representation for modeling 3D surface. As shown in Fig. 2(a), 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 [32] conditions it on pixel-aligned image features using the following formulation:

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

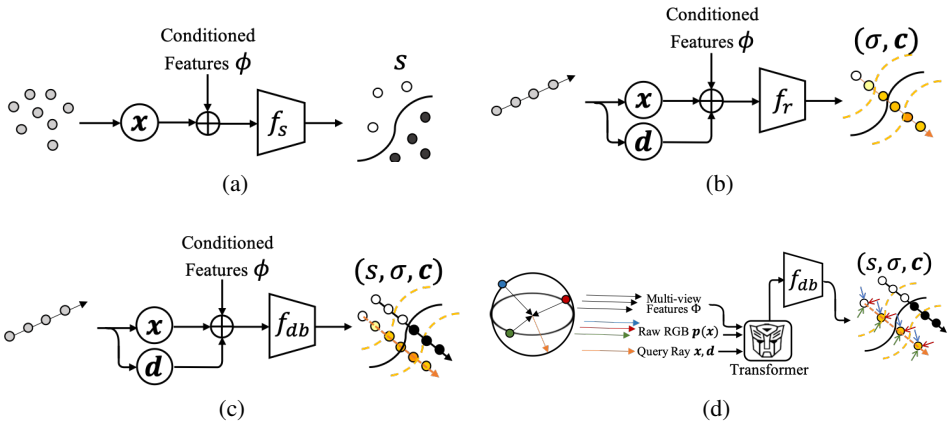


Figure 2: Comparison of different neural field representations. **(a)** Neural surface field in PIFu [32]. **(b)** Neural radiance field in PixelNeRF [45]. **(c)** The proposed DoubleField. **(d)** DoubleField with the raw ultra-high-resolution inputs.

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 the geometry and appearance, it isolates geometry and texture and makes the learning space of texture discontinuous and highly concentrated around the surface. Such a discontinuous texture space hinders the optimization process under the texture supervisions using differentiable rendering techniques [27].

Neural Radiance Field As shown in Fig. 2(b), NeRF [26] 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, a 2D image of novel view can be rendered by the integration along camera 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 [45] 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, inference, and finetuning of its derivative solutions [45, 29] time-consuming.

4 DoubleField Representation

In our approach, a novel neural field representation DoubleField is proposed to bridge the surface field and radiance field. As shown in Fig. 2(c), our DoubleField can be formulated as a mutual implicit function f_{db} represented by multi-layer perceptrons (MLPs) to fit both the surface field and the radiance field: $f_{db}(\boldsymbol{x}, \boldsymbol{d}) = (s, \sigma, \boldsymbol{c})$. The MLP shared between two fields imposes a geometry constraint to the radiance field in an implicit manner and encourages a more consistent density distribution for neural rendering.

Network Architecture DoubleField is composed of a shared MLP for double embedding and two individual MLPs for geometry and texture modeling. Without loss of generality, DoubleField is also conditioned on pixel-aligned images features $\phi(\boldsymbol{x}, \boldsymbol{I})$. In our implementation, given the query point $\boldsymbol{x}$ and viewing direction $\boldsymbol{d}$, a *double MLP* E_{db} learns a shared double embedding $\mathbf{e}_{db}$, which is further decoded by two MLPs E_g and E_c for the prediction of the geometry and the texture fields. Overall, the DoubleField conditioned on pixel-aligned features can be written as:

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

where $\gamma(\boldsymbol{x})$ is the positional encoding of $\boldsymbol{x}$, E_g is a *geometry MLP* for the prediction of occupancy in the surface field and the density in the radiance field, while E_c is a *texture MLP* for prediction of the color in the radiance field. Both the MLPs E_g and E_c are much lighter than the *double MLP* E_{db}, which implicitly builds a strong association between the two fields and imposes the surface constraint on the learning of the radiance field.

Sampling Strategy To facilitate the learning process, we make full use of the surface field and propose a surface-guided sampling strategy for DoubleField. As illustrated in Fig.3, 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 the camera parameters of the rendering view and the camera rays $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, a uniform sampling is applied along the ray in the depth bounds $[t_n, t_f]$ with N_s sampling points. We query the surface field value of each point to determine the first intersection position on the surface. These intersection positions are used to guide the sampling at a more fine-grained level by considering the radiance field surrounding the intersected surface in a interval of δ with N_r sampling points. More details about the sampling process can be found in the Supp.Mat.

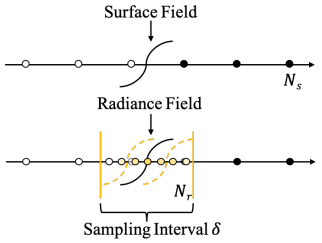


Figure 3: Illustration of the surface-guided sampling strategy.

4.1 DoubleField with Multi-view Inputs

DoubleField can be easily extended to multi-view settings, where pixel-aligned features will be extracted from the multi-view images and then fused together for the prediction of the two fields. Specifically, given image inputs $\{\boldsymbol{I}^i\}_1^n$ 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 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 viewpoint, and $\psi(...)$ is a feature fusion operation such as average pooling [32] or self-attention [49]. 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: $f_{db}(\boldsymbol{x}, \boldsymbol{d}|\boldsymbol{\Phi}(\boldsymbol{x})) = (s, \sigma, \boldsymbol{c})$.

4.2 DoubleField on Ultra-high-resolution Domain

The merits of DoubleField also pave the way to exploit solutions for higher-quality neural rendering. As discussed in previous Sections, the learning on coarse image features limits the quality of the final rendering results. To overcome this issue, we further augment DoubleField to take the images at the original resolution as additional condition inputs (see Fig. 2(d)):

$$f_{db}(\boldsymbol{x}, \boldsymbol{d}|\boldsymbol{\Phi}(\boldsymbol{x}), \boldsymbol{p}(\boldsymbol{x})) = (s, \sigma, \boldsymbol{c}), \tag{6}$$

where $\boldsymbol{p}(\boldsymbol{x})$ denotes the pixel RGB values at the projection of $\boldsymbol{x}$.

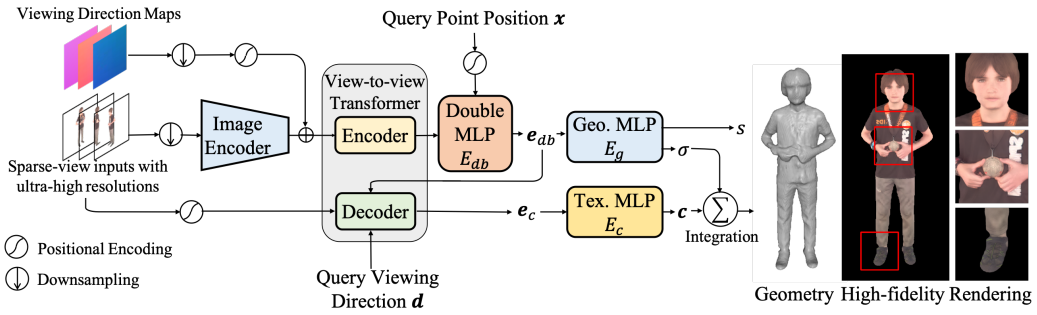


Figure 4: High-fidelity neural human rendering based on DoubleField.

Motivated by the text-to-text problem in NLP [31], we design a view-to-view transformer to learn geometry and appearance on the ultra-high-resolution domain with sparse-view inputs. The view-to-view transformer fuses the raw RGB values and multi-view features in its encoder and produce features at the novel-view space by its decoder. Moreover, the raw RGB values are mapped to a higher dimensional space as the colored encoding for the learning of high-frequency appearance variation. The key components of the view-to-view transformer are presented as follows.

Colored Encoding Similar to the positional encoding [36, 26], the raw RGB values $\boldsymbol{p}$ on each pixel of an ultra-high-resolution image $\mathbf{I}$ are embedded as a colored encoding $\gamma(\boldsymbol{p})$ using the sine and cosine functions of different frequencies [26]. In this way, each single RGB value is mapped to a higher dimensional space, which significantly improves the performance and accelerates the convergence speed. This is consistent with the previous work on neural field representations [26] and NLP [36].

Encoder In our view-to-view transformer, a "fully-visible" attention mask is used in the encoder, which encourages the model attending to each view related to the novel view by the self-attention mechanism. The encoder acts as the feature fusion operation ψ in Eq. 5 to obtain the fused features $\mathbf{\Phi}$, which will be fed into the *double MLP* E_{db} for the generation of the double embedding $\boldsymbol{e}_{db} = E_{db}(\gamma(\boldsymbol{x}), \mathbf{\Phi})$.

Decoder The decoder of our view-to-view transformer maps the features learned on sparse-view inputs to the features of novel viewpoints. Specifically, the decoder takes the double embedding $\boldsymbol{e}_{db}$, the query viewing direction $\boldsymbol{d}$, the positional encoding of the query point $\boldsymbol{x}$, and the colored encoding of the RGB values $\boldsymbol{p}(\boldsymbol{x})$ as inputs to obtain texture embedding $\boldsymbol{e}_c$:

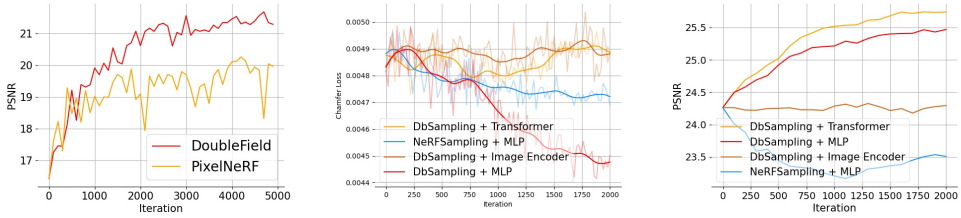
$$\boldsymbol{e}_c = D_{v2v}(\boldsymbol{e}_{db}, \gamma(\boldsymbol{p}(\boldsymbol{x}))|\boldsymbol{d}). \tag{7}$$

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

5 Learning High-fidelity Human Rendering

The efficacy of our DoubleField is validated on the geometry and appearance reconstruction from sparse-view human images. As illustrated in Fig. 4, given sparse multi-view images and the corresponding ray directions, the encoder of our view-to-view transformer serves as the operation ψ to fuse low-resolution image features from different viewpoints and output the fused features using Eq. 5. 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 viewing direction $\boldsymbol{d}$, 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 $\boldsymbol{c}$.

Though our network can be directly trained on ultra-high-resolution images, the expensive training time cost on such a high-fidelity domain is still a problem. For a more feasible solution, we adopt a transductive transfer learning scheme to divide the problem into two phases: low-resolution pre-training and high-fidelity finetuning. In the pre-training phase, the network learns two coarse prior on down-sampling images: 1) A general geometry and appearance prior of human. 2) A fusion prior of



(a) Overfitting (b) Finetuning geometry (c) Finetuning color

Figure 5: (a) Overfitting experiments. (b)(c) The curves of Chamfer distance and PSNR under different finetuning strategies. Curves come from the experiments of one randomly-selected model.

multi-view features and raw RGB values. Specifically, to train our model, we collect human models from 3D scan dataset such as Twindom[1] and render low-resolution images with the size of 512×512. In the finetuning phase, the network takes the ultra-high-resolution images from sparse multi-view of specific human as inputs and is finetuned using multi-view self-supervision. In this way, the model pre-trained on low-resolution images is adapted to the ultra-high-resolution domain. More details about the transfer learning scheme can be found in the Supp.Mat.

6 Experiment

We evaluate our DoubleField representation and view-to-view transformer on several dataset: 1. Twindom dataset, we split 1,700 human models into 1,500 models for training and 200 models for evaluation. 2. THuman2.0 dataset [46], which is a publicly-available dataset consisting of 500 high-quality human models. We first validate the proposed DoubleField from the aspects of training, inference, and finetuning. Then we present experimental results of various strategies on the finetuning of the view-to-view transformer. Finally, we compare our solutions with prior state-of-the-art approaches.

6.1 Efficiency of DoubleField

To validate the efficiency of DoubleField representation, we first evaluate the training process via an overfitting experiment and compare with PixelNeRF [45]. We randomly select one model from the Twindom dataset and render 60 views for supervision. During training, we use a fixed single-view as input and a network with the basic DoubleField in Eq. 4 (i.e., no multi-view fusion modules and transformer). For the loss function, PixelNeRF is trained using with the sampling strategy in NeRF [26] to formulate rendering loss with other 59 views, while DoubleField is trained using the sampling strategy in PIFu [32] for the learning of the surface field and the proposed surface-guided sampling strategy for the learning of the radiance field. We evaluate the performances using another fixed viewpoint image for every 100 iterations during the training and report results in Fig. 5. We can see that our method achieves fast convergence while PixelNeRF is struggling to reconstruct the entangled appearance and geometry, which proves the imposed geometry constraint of surface field in an implicit form is helpful to the training process.

Based on the proposed DoubleField representation and the surface-guided sampling strategy, our network not only achieves much faster rendering speed but also acquires much less memory than NeRF. The reasons are: 1) The number of sampling points (set to 16 in our experiments) around the surface can be much less than the importance sampling in NeRF. 2) During training, the determination of the intersection points on the surface has no gradient for back-propagation so that the surface-guided sampling saves both time and memory. 3) A coarse mesh can be directly extracted from the surface field using marching cube at inference, which greatly reduces the total number of query rays by removing background and allows fast intersection detection using the depth map. The comparison of different sampling strategies can be found in the Supp.Mat.

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

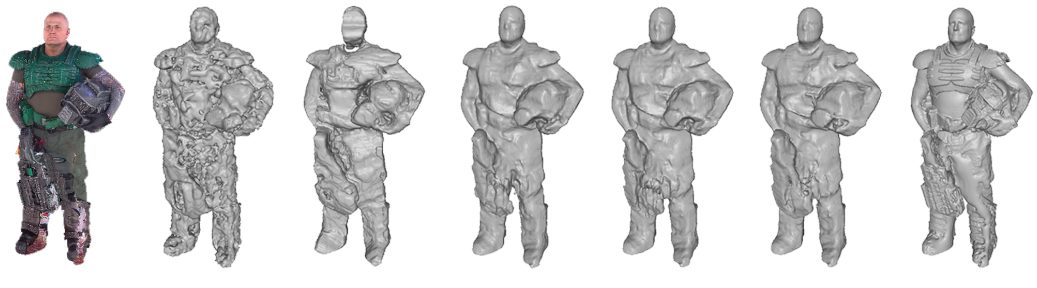


Figure 6: Comparisons of geometry reconstruction results.

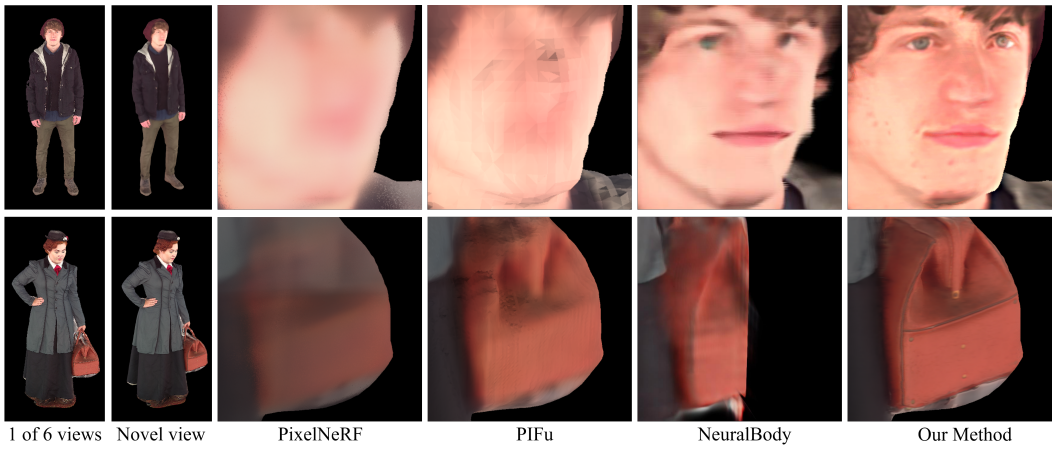


Figure 7: Comparisons of appearance reconstruction results on Twindom dataset. PixelNeRF and our method are finetuned with additional 4,000 iterations.

6.2 Finetuning with Different Strategies

Our network consists of several modules for image feature extraction, fusion, and final DoubleField prediction, which means there are various possible strategies for finetuning. Directly finetuning the whole the network with sparse multiview supervisions is prone to overfitting. To overcome this issue, we conduct experiments on 10 models selected randomly from the Twindom test dataset to figure out a plausible strategy that can avoid overfitting with fast convergence. For the finetuning on each model, there are sparse-view images from 6 fixed viewpoints in total, and we randomly pick 4 images as the network input and 1 image for supervision in each iteration.

In the finetuning process, we evaluate the geometry reconstruction using chamfer distance for every 20 iterations and the rendering results using PSNR metric with 24 novel fixed viewpoints for every 100 iterations. We adopt different finetuning strategy including: 1) Finetuning on the image encoder. 2) Finetuning on the view-to-view transformer. 3) Finetuning on the MLPs (the double MLP, the geometry MLP and the texture MLP). In each finetuning process, we fix other modules and finetune only the specific module. We also conducted finetuning on the MLPs using the NeRF sampling strategy. The results are reported in Fig. 5, which shows that finetuning transformer can help to achieve higher fidelity rendering and finetuning downstreaming MLPs can refine both geometry and color reconstruction. Besides, using the NeRF sampling strategy instead of the proposed surface-guided sampling strategy leads to deteriorated performance due to overfitting, which further demonstrates that our sampling strategy alleviates the overfitting issue and contributes to better reconstruction.

In summary, we adopt a bottom-up strategy to finetune our network for high resolution inputs. Specifically, we first finetune MLPs to refine geometry and color, and then finetune the transformer and the texture MLP to achieve high-fidelity rendering.

Method	Twindom (6 views Col.)		THuman2.0 (6 views Col.)		Twindom (6 views Geo.)		THuman2.0 (6 views Geo.)	
	PSNR	SSIM	PSNR	SSIM	Chamfer	P2S	Chamfer	P2S
PIFu [32]	20.80	0.805	22.35	0.846	0.754	0.716	0.710	0.613
PIFu+DVR	20.65	0.804	22.17	0.843	**0.746**	**0.701**	0.709	**0.611**
PixelNeRF [45]	21.57	0.808	22.95	0.854	0.945	0.931	0.815	0.725
Our Method	**22.95**	**0.842**	**24.23**	**0.880**	0.750	0.707	**0.708**	0.612
NeuralBody [29]	20.69	0.808	22.65	0.862	1.597	2.146	1.528	2.126
PIFu+DVR (Ft)	21.62	0.812	23.08	0.855	0.779	0.736	0.724	0.623
PixelNeRF (Ft)	21.85	0.813	23.57	0.863	1.072	1.052	0.790	0.701
Our Method (Ft)	**23.56**	**0.857**	**25.10**	**0.905**	**0.721**	**0.694**	**0.673**	**0.601**

Table 2: Quantitative results on Twindom dataset and THuman2.0 dataset for human geometry and appearance reconstruction. Ft denotes the approaches finetuned with 4,000 iterations.

Figure 8: Comparisons with NeuralBody. 4 images on the left are from ZJU-mocap, and 4 images on the right are from real world multi-view (5 views) system. Each video has 300 frames and we train NeuralBody for 20 hours.

6.3 Comparisons with State-of-the-art Approaches

In this section, we compare DoubleField with the state-of-the-art approaches built upon the surface field and radiance field, including PIFu [32], PixelNeRF [45], and NeuralBody [29]. We also implement DVR [27] 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 replace the average pooling operation in PIFu [32] and PixelNeRF [45] with self-attention modules for multi-view feature fusion and retrain their networks with the same training settings and datasets.

Geometry Reconstruction For the comparison with NeuralBody [29], we regard NeuralBody as a frame-based method and train it on 6 viewpoint inputs for 15 hours. We quantitatively evaluate the geometry recovery performance using the point-to-surface distance and Chamfer distance and report results in Table. 2. We can see that our method without finetuning achieves competitive results compared with PIFu and PIFu+DVR. With finetuning, our method can improve the quality of geometry based on the DoubleField representation. Qualitative results are illustrated in Fig. 6. Unlike PixelNeRF and NeuralBody, 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 two fields.

Appearance Reconstruction To evaluate different methods on the appearance reconstruction, we prepare images of 4K resolution rendered from 30 viewpoints. We 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 DoubleField representation, our method can achieve high-fidelity rendering. Under the transfer learning scheme, 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 for scenarios like object interactions and loose clothes (such as long skirts) as shown in Fig. 7.

Comparisons on Real-world Data To demonstrate the robustness of our method, we also evaluate our method using the ZJU-mocap dataset and real world captured multi-view videos. The results are shown in Fig. 8 and Fig. 9. Our method produces comparable rendering results on the ZJU-mocap dataset but using much less time for network finetuning (<20 minutes V.S. >10 hours). Moreover,

Figure 9: Comparisons on real world data.

our method does not rely on human shape prior SMPL [22] like NeuralBody and achieves higher quality results 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.

7 Discussion and Future Works

We propose the DoubleField representation to combine the merits of geometry and appearance fields for high-fidelity human rendering. Though our approach achieves superior performances on the reconstruction from sparse-view ultra-high-resolution inputs, the high-quality 3D human models are still essential to learn the geometry prior. Moreover, the effort of finetuning on the geometry refinement is limited, which hinders our approach to handle extremely challenging poses.

In our work, the associations between two fields are built in an implicit manner. A more unified and explicit formulation is still deserved to be exploited. Besides, the proposed view-to-view transformer and the transfer learning scheme provide novel solutions for high-fidelity rendering. We hope our approach can enlighten the follow-up work in the field of free-viewpoint human rendering.

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

[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.

[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.

[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.

[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.

[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.

[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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

[20] 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.

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

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.
[30] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021.
[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2019.
[32] 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.
[33] 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.
[34] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3D-aware image synthesis. In *NeurIPS*, volume 33, 2020.
[35] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019.
[36] 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.
[37] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *TOG*, pages 1–9, 2008.
[38] 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.
[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.
[40] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *CVPR*, pages 1682–1691, 2020.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021.
[46] 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.
[47] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.
[48] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
[49] 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. *arXiv e-prints*, pages arXiv–2105, 2021.
[50] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *TPAMI*, 2021.