This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

GM-NeRF: Learning Generalizable Model-based Neural Radiance Fields from Multi-view Images

Jianchuan Chen[1*] Wentao Yi[1*] Liqian Ma[2†] Xu Jia[1] Huchuan Lu[1†]
[1] Dalian University of Technology, China [2] ZMO AI Inc.

Abstract

In this work, we focus on synthesizing high-fidelity novel view images for arbitrary human performers, given a set of sparse multi-view images. It is a challenging task due to the large variation among articulated body poses and heavy self-occlusions. To alleviate this, we introduce an effective generalizable framework Generalizable Model-based Neural Radiance Fields (GM-NeRF) to synthesize free-viewpoint images. Specifically, we propose a geometry-guided attention mechanism to register the appearance code from multi-view 2D images to a geometry proxy which can alleviate the misalignment between inaccurate geometry prior and pixel space. On top of that, we further conduct neural rendering and partial gradient backpropagation for efficient perceptual supervision and improvement of the perceptual quality of synthesis. To evaluate our method, we conduct experiments on synthesized datasets THuman2.0 and Multi-garment, and real-world datasets Genebody and ZJUMocap. The results demonstrate that our approach outperforms state-of-the-art methods in terms of novel view synthesis and geometric reconstruction.

1. Introduction

3D digital human reconstruction has a wide range of applications in movie production, telepresence, 3D immersive communication, and AR/VR games. Traditional digital human production relies on dense camera arrays [10, 14] or depth sensors [12, 20] followed by complex graphics rendering pipelines for high-quality 3D reconstruction, which limits the availability to the general public.

Reconstructing 3D humans from 2D images captured by sparse RGB cameras is very attractive due to its low cost and convenience. This field has been studied for decades [21, 46, 50]. However, reconstruction from sparse RGB cameras is still quite challenging because of: 1) heavy self-occlusions of the articulated human body; 2) inconsistent lighting and sensor parameters between different cameras; 3) highly non-rigid and diverse clothes.

Figure 1. **The effect of inaccurately estimated SMPL**. Compared with GNR [8] and KeypointNeRF [26], our method still yields a reasonable result.

In recent years, with the rise of learning-based methods, we can reconstruct high-quality digital humans from sparse cameras. Learning-based methods [32, 36, 43, 49, 52] have made great processes, however, they lack multi-view geometric consistency due to the mere usage of a 2D neural rendering network. To address this problem, many recent works [5, 47, 54] adopt neural radiance fields as 3D representations, which achieves outstanding performance on novel view synthesis. However, these methods are not robust to unseen poses without the guidance of human geometric prior.

To better generalize to unseen poses, NeuralBody [31] introduces a statistical body model SMPL [23] into neural radiance fields which can reconstruct vivid digital humans from a sparse multi-view video. However, NeuralBody is designed for identity-specific scenarios, which means it requires laborious data collection and long training to obtain the model for one person. Such a limitation restricts its application in general real-world scenarios.

In this work, we focus on synthesizing high-fidelity novel view images for arbitrary human performers from a set of sparse multi-view images. Towards this goal, some very recent works [7, 8, 19, 26] propose to aggregate multi-view pixel-aligned features using SMPL as a geometric prior. However, these methods usually assume perfect geometry (*e.g.* accurate SMPL [23] estimation from 2D images) which is not applicable in practical applications. In

*Equal contribution. †Corresponding authors. Codes are available at https://github.com/JanaldoChen/GM-NeRF

practice, the geometry error does affect the reconstruction performance significantly. As illustrated in the red box of Fig. 1, when the estimated SMPL does not align well with RGB image, prior SMPL-dependent methods [8, 26] yield blurry and distorted results. The such performance gap is caused by the misalignment between the 3d geometry (*i.e.* SMPL) and the pixel space (*i.e.* pixel-aligned feature and ground-truth image). Specifically, the misalignment will cause: 1) blur and distortion when fusing the geometry and pixel-aligned features; 2) unsuitable supervision during training with a pixel-wise loss like L1 or L2. To alleviate the issue of misalignment, we propose to take the geometry code as a proxy and then register the appearance code onto the geometry through a novel geometry-guided attention mechanism. Furthermore, we leverage perceptual loss to reduce the influence of misalignment and promote sharp image synthesis, which is evaluated at a higher level with a larger perceptual field. It is non-trivial to apply perceptual loss in NeRF-based methods as the perceptual loss requires a large patch size as input which is memory-consuming through volume rendering. We introduce 2D neural rendering and partial gradient backpropagation to alleviate the memory requirement and enhance the perceptual quality.

To summarize, our work contributes as follows:

- A novel generalizable model-based framework GM-NeRF is proposed for the free-viewpoint synthesis of arbitrary performers.
- To alleviate the misalignment between 3D geometry and the pixel space, we propose geometry-guided attention to aggregate multi-view appearance and geometry proxy.
- To enable perceptual loss supervision to further alleviate misalignment issues, we adopt several efficient designs including 2D neural rendering and partial gradient backpropagation.

2. Related work

Implict Neural Representation. Implicit neural representations (also known as coordinate-based representations) are a popular way to parameterize content of all kinds, such as audio, images, video, or 3D scenes [27, 38, 40, 42]. Recent works [25, 27, 28, 40] build neural implicit fields for geometric reconstruction and novel view synthesis achieving outstanding performance. The implicit neural representation is continuous, resolution-independent, and expressive, and is capable of reconstructing geometric surface details and rendering photo-realistic images. While explicit representations like point clouds [1, 52], meshes [43], and voxel grids [22, 25, 39, 44] are usually limited in resolution due to memory and topology restrictions. One of the most popular implicit representations - Neural Radiance Field (NeRF) [27] - proposes to combine the neural radiance field with differentiable volume for photo-realistic novel views rendering of static scenes. However, NeRF requires optimizing the 5D neural radiance field for each scene individually, which usually takes hours to converge. Recent works [5, 47, 54] try to extend NeRF to generalization with sparse input views. In this work, we extend the neural radiance field to a general human reconstruction scenario by introducing conditional geometric code and appearance code.

3D Model-based Human Reconstruction With the emergence of human parametric models like SMPL [23, 29] and SCAPE [3], many model-based 3D human reconstruction works have attracted wide attention from academics. Benefiting from the statistical human prior, some works [2, 4, 9, 18] can reconstruct the rough geometry from a single image or video. However, limited by the low resolution and fixed topology of statistical models, these methods cannot represent arbitrary body geometry, such as clothing, hair, and other details well. To address this problem, some works [33,34] propose to use pixel-aligned features together with neural implicit fields to represent the 3D human body, but still have poor generalization for unseen poses. To alleviate such generalization issues, [15, 35, 57] incorporate the human statistical model SMPL [23, 29] into the implicit neural field as a geometric prior, which improves the performance on unseen poses. Although these methods have achieved stunning performance on human reconstruction, high-quality 3D scanned meshes are required as supervision, which is expensive to acquire in real scenarios. Therefore, prior works [15, 33, 34, 57] are usually trained on synthetic datasets and have poor generalizability to real scenarios due to domain gaps. To alleviate this limitation, some works [6, 30, 31, 41, 48, 51] combine neural radiance fields [27] with SMPL [23] to represent the human body, which can be rendered to 2D images by differentiable rendering. Currently, some works [7, 8, 19, 26, 35, 53] can quickly create neural human radiance fields from sparse multi-view images without optimization from scratch. While these methods usually rely on accurate SMPL estimation which is not always applicable in practical applications.

3. Method

We introduce an effective framework GM-NeRF for novel view synthesis and 3D human reconstruction as illustrated in Fig. 2. GM-NeRF learns generalizable model-based neural radiance fields from calibrated multi-view images by introducing a parametric model SMPL as a geometric prior, which can generalize to unseen identity and unseen pose.

Given m calibrated multi-view images $\{I_k\}_{k=1}^{m}$ of a person, we use Easymocap [11] to obtain the SMPL [23] parameters $\mathbf{M}(\theta, \beta)$ of the person. We feed the multi-view images into the encoder network $\mathbf{E}$ to extract multi-view feature maps,

$$H_k = \mathbf{E}(I_k), \quad k = 1, 2, \ldots, m. \tag{1}$$

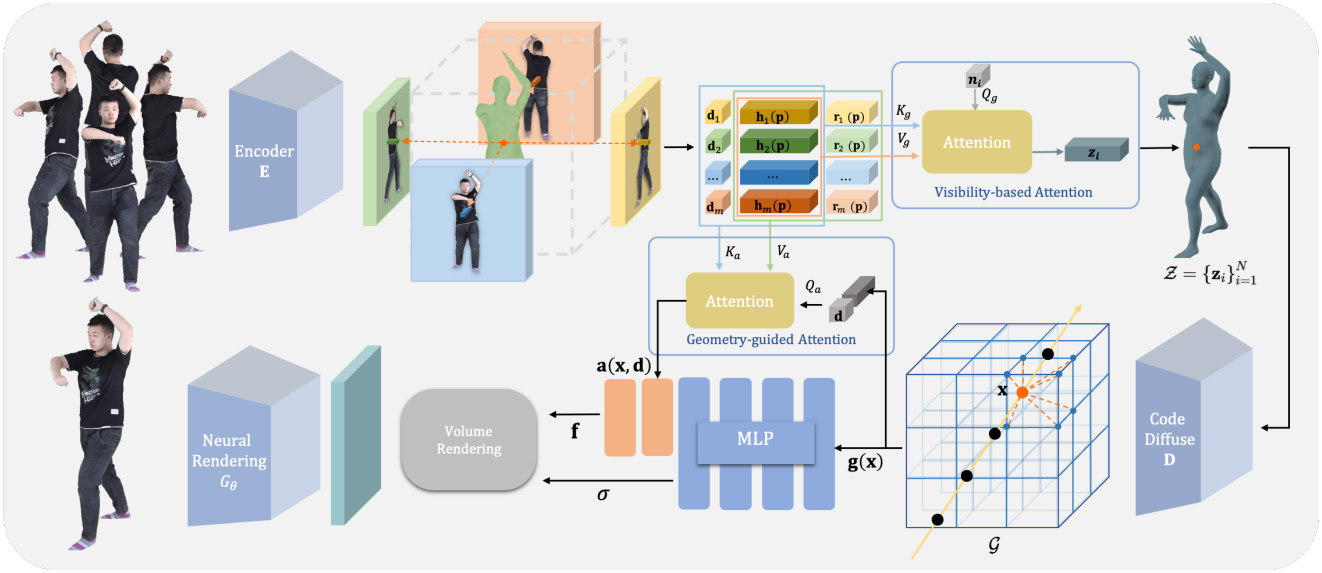

Figure 2. **The architecture of our method**. Given m calibrated multi-view images and registered SMPL, we build the generalizable model-based neural human radiance field. First, we utilize the image encoder to extract multi-view image features, which are used to provide geometric and appearance information, respectively. In order to adequately exploit the geometric prior, we propose the visibility-based attention mechanism to construct a structured geometric body embedding, which is further diffused to form a geometric feature volume. For any spatial point $\mathbf{x}$, we trilinearly interpolate the feature volume $\mathcal{G}$ to obtain the geometric code $\mathbf{g}(\mathbf{x})$. In addition, we also propose geometry-guided attention to obtain the appearance code $\mathbf{a}(\mathbf{x}, \mathbf{d})$ directly from the multi-view image features. We then feed the geometric code $\mathbf{g}(\mathbf{x})$ and appearance code $\mathbf{a}(\mathbf{x}, \mathbf{d})$ into the MLP network to build the neural feature field $(\mathbf{f}, \sigma) = F(\mathbf{g}(\mathbf{x}), \mathbf{a}(\mathbf{x}, \mathbf{d}))$. Finally, we employ volume rendering and neural rendering to generate the novel view image.

For any 3D position $\mathbf{p}$, we can project it onto the feature map H_k according to the corresponding camera parameters, which is defined as $\pi_k(\cdot)$, then use bilinear interpolation $\Psi(\cdot)$ to obtain the pixel-aligned feature $\mathbf{h}_k(\mathbf{p})$ and pixel-aligned color $\mathbf{r}_k(\mathbf{p})$ as follows,

$$\begin{aligned} \mathbf{h}_k(\mathbf{p}) &= \Psi(H_k, \pi_k(\mathbf{p})), \\ \mathbf{r}_k(\mathbf{p}) &= \Psi(I_k, \pi_k(\mathbf{p})). \end{aligned} \tag{2}$$

In order to adequately exploit the geometric prior, we propose the visibility-based attention mechanism to construct a structured geometric body embedding, which is further diffused to form a geometric feature volume (Sec. 3.1). Afterward, we trilinear interpolate each spatial point $\mathbf{x}$ in the feature volume $\mathcal{G}$ to obtain the geometric code $\mathbf{g}(\mathbf{x})$. To avoid the misalignment between the appearance code and geometry code, we utilize the geometry code as a proxy and then register the appearance code $\mathbf{a}(\mathbf{x}, \mathbf{d})$ directly from the multi-view image features with a novel geometry-guided attention mechanism (Sec. 3.2). We then feed the geometric code $\mathbf{g}(\mathbf{x})$ and appearance code $\mathbf{a}(\mathbf{x}, \mathbf{d})$ into the MLP network to build the neural feature field $(\mathbf{f}, \sigma) = F(\mathbf{g}(\mathbf{x}), \mathbf{a}(\mathbf{x}, \mathbf{d}))$ followed by volume rendering and neural rendering for novel view image generation (Sec. 3.3). To obtain high-quality results, we carefully design an optimization objective including a novel normal regularization. (Sec. 3.4) as well as an efficient training strategy (Sec. 3.5).

3.1. Structured Geometric Body Embedding

Different from neural radiance fields on general scenes, we introduce a parametric body model to provide the geometric prior for constructing the neural human radiance field, which can enhance generalizability under unseen poses. In our experiments, we choose the SMPL [23] model as the parametric model. The SMPL [23] model $\mathbf{M}(\theta, \beta)$ is a mesh with $N = 6,890$ vertices $\{\mathbf{v}_i\}_{i=1}^N$, where it is mainly controlled by the pose parameter θ, and the shape parameter β. NeuralBody [31] optimizes a set of structured latent codes from scratch on vertices of the SMPL model for each specific identity. However, not only does it fail to represent a new identity but also has poor generalizability on unseen poses. To address such limitation, we extract the structured latent codes $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$ from the multi-view feature map H_k as a geometric embedding to represent arbitrary identities. For vertex $\mathbf{v}_i$, we design a visibility-based attention mechanism as shown in Fig. 2 to fuse multi-view features.

$$\begin{aligned} \mathbf{Q}_g(\boldsymbol{v}_i) &= F_Q^g(\boldsymbol{n}_i) \\ \mathbf{K}_g(\boldsymbol{v}_i) &= F_K^g(\{\mathbf{h}_k(\boldsymbol{v}_i) \oplus \mathbf{d}_k\}_{k=1}^m) \\ \mathbf{V}_g(\boldsymbol{v}_i) &= F_V^g(\{\mathbf{h}_k(\boldsymbol{v}_i)\}_{k=1}^m) \\ \mathbf{z}_i &= F^g\left(Att\left(\mathbf{Q}_g(\boldsymbol{v}_i), \mathbf{K}_g(\boldsymbol{v}_i), \mathbf{V}_g(\boldsymbol{v}_i)\right)\right) \end{aligned} \tag{3}$$

where $\oplus$ is the concatenation operator, and $\boldsymbol{n}_i$ is the normal of the vertex $\boldsymbol{v}_i$. F_Q^g, F_K^g, F_V^g denote the geometric linear

layers producing the query, key, and value matrices $\mathbf{Q}_g(\boldsymbol{v}_i)$, $\mathbf{K}_g(\boldsymbol{v}_i)$, $\mathbf{V}_g(\boldsymbol{v}_i)$, respectively. Att is the attention mechanism proposed by [45]. F^g is the geometric feed-forward layer. The intuition of this visibility-based attention mechanism is that the closer the input camera direction $\mathbf{d}_k$ is to the normal $\boldsymbol{n}_i$, the more the corresponding feature contributes. As shown in Fig. 5, the visualization result demonstrates the plausibility of this design.

Similar to NeuralBody [31], we use SparseConvNet [13] $\mathbf{D}$ to diffuse the structured latent codes $\{\mathbf{z}_i\}_{i=1}^N$ into the nearby space to form a 3D feature volume $\mathcal{G}$.

$$\begin{aligned} \mathcal{G} &= \mathbf{D}(\{\mathbf{z}_i\}_{i=1}^N) \\ \mathbf{g}(\mathbf{x}) &= \Phi(\mathbf{x}, \mathcal{G}) \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ is the trilinear interpolation operation, which is applied to obtain the geometric code $\mathbf{g}(\mathbf{x})$ for any 3D position $\mathbf{x}$ during volume rendering.

3.2. Multi-View Appearance Blending

Although the structured geometric body embedding provides a robust geometric prior, high-frequency appearance details such as wrinkles and patterns are lost, due to the low resolution and the minimally-clothed topology of the parametric model. In practice, inaccurate SMPL estimation will lead to the misalignment between the 3D geometry and pixel space, which will cause blur and distortion when fusing the geometry and pixel-aligned feature. To solve this problem, we design a geometry-guided attention mechanism as shown in Fig. 2, which utilizes the geometry code as a proxy and then registers the appearance code $\mathbf{a}(\mathbf{x}, \mathbf{d})$ directly from the multi-view image features for any 3D position $\mathbf{x}$ and view direction $\mathbf{d}$.

$$\begin{aligned} \mathbf{Q}_a(\mathbf{x}) &= F_Q^a\left(\mathbf{g}(\mathbf{x}) \oplus \mathbf{d}\right) \\ \mathbf{K}_a(\mathbf{x}) &= F_K^a(\{\mathbf{h}_k(\mathbf{x}) \oplus \mathbf{d}_k\}_{k=1}^m) \\ \mathbf{V}_a(\mathbf{x}) &= F_V^a(\{\mathbf{h}_k(\mathbf{x}) \oplus \mathbf{r}_k(\mathbf{x})\}_{k=1}^m) \\ \mathbf{a}(\mathbf{x}, \mathbf{d}) &= F^a\left(Att\left(\mathbf{Q}_a(\mathbf{x}), \mathbf{K}_a(\mathbf{x}), \mathbf{V}_a(\mathbf{x})\right)\right) \end{aligned} \tag{5}$$

where F_Q^a, F_K^a, F_V^a denote the appearance layers producing the query, key, and value matrices $\mathbf{Q}_a(\mathbf{x})$, $\mathbf{K}_a(\mathbf{x})$, $\mathbf{V}_a(\mathbf{x})$, respectively. F^a is the appearance feed-forward layer.

3.3. Differential Rendering

After we get the geometric code and appearance code of any 3D point, we design a two-stage MLP network $F(\cdot)$ to build the neural feature field.

$$(\mathbf{f}, \sigma) = F(\mathbf{g}(\mathbf{x}), \mathbf{a}(\mathbf{x}, \mathbf{d})) \tag{6}$$

Unlike classical NeRF [27], which regresses color $\mathbf{c}$ and density σ, our decoder outputs the intermediate feature $\mathbf{f}$ and density σ. However, the original volume rendering process is memory-consuming, we use a combination of volume rendering and neural rendering to get the final image.

3D Volume Rendering. We use the same volume rendering techniques as in NeRF [27] to render the neural radiance field into a 2D image. Then the pixel colors are obtained by accumulating the colors and densities along the corresponding camera ray τ. In practice, the continuous integration is approximated by summation over sampled N points $\{\mathbf{x}_i\}_{i=1}^N$ between the near plane and the far plane along the camera ray τ.

$$\begin{aligned} \mathcal{F}(\tau) &= \sum_{i=1}^N \alpha_i\left(\mathbf{x}_i\right) \prod_{j<i}\left(1-\alpha_j\left(\mathbf{x}_j\right)\right) \mathbf{f}\left(\mathbf{x}_i\right) \\ \mathcal{M}(\tau) &= \sum_{i=1}^N \alpha_i\left(\mathbf{x}_i\right) \prod_{j<i}\left(1-\alpha_j\left(\mathbf{x}_j\right)\right) \\ \alpha_i(\mathbf{x}) &= 1-\exp\left(-\sigma(\mathbf{x})\delta_i\right) \end{aligned} \tag{7}$$

where $\delta_i = \left\|\mathbf{x}_{i+1} - \mathbf{x}_i\right\|_2$ is the distance between adjacent sampling points. $\alpha_i(\mathbf{x})$ is the alpha value for $\mathbf{x}$. The intermediate feature image $I_{\mathcal{F}} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times M_{\mathcal{F}}}$ and the silhouette image $I_{\mathcal{M}} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 1}$ is obtained by Eq. (7).

2D Neural Rendering. We utilize a 2D convolutional network G_θ to convert the intermediate feature image $I_{\mathcal{F}} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times M_{\mathcal{F}}}$ rendered by volume rendering into the final synthesized image $I_t \in \mathbb{R}^{H \times W \times 3}$.

$$I_t \longleftarrow G_\theta\left(I_{\mathcal{F}}\right) \tag{8}$$

where θ is the parameters of the 2D neural rendering network G, which means the rendering procedure is learnable.

3.4. Loss Functions

To stabilize the training procedure, we adopt the pixel-wise L2 loss widely used in [8, 19, 54] to constrain the rendered image I_t and the alpha image $I_{\mathcal{M}}$.

$$\mathcal{L} = \lambda_r \left\|\tilde{I}_t - I_t\right\|_2^2 + \lambda_s \left\|\tilde{I}_{\mathcal{M}} - I_{\mathcal{M}}\right\|_2^2 \tag{9}$$

where $\tilde{I}_t$, $\tilde{I}_{\mathcal{M}}$ are the ground-truth of the RGB image and silhouette image, respectively and λ_r, λ_s are the weights. Beyond that, we also introduce the following loss functions to optimize the networks together,

Perceptual Loss . We use a perceptual loss [16] based on the VGG Network [37]. It is more effective when the size of the images is closer to the network input, while it is memory intensive to render the whole image by volume rendering. To address these limitations, we adapt both neural rendering as well as partial gradient backpropagation.

$$\mathcal{L}_p = \sum \frac{1}{N^j}\left|p^j\left(\tilde{I}_t\right) - p^j\left(I_t\right)\right| \tag{10}$$

where p^j is the activation function and N^j is the number of elements of the j-th layer in the pretrained VGG network.

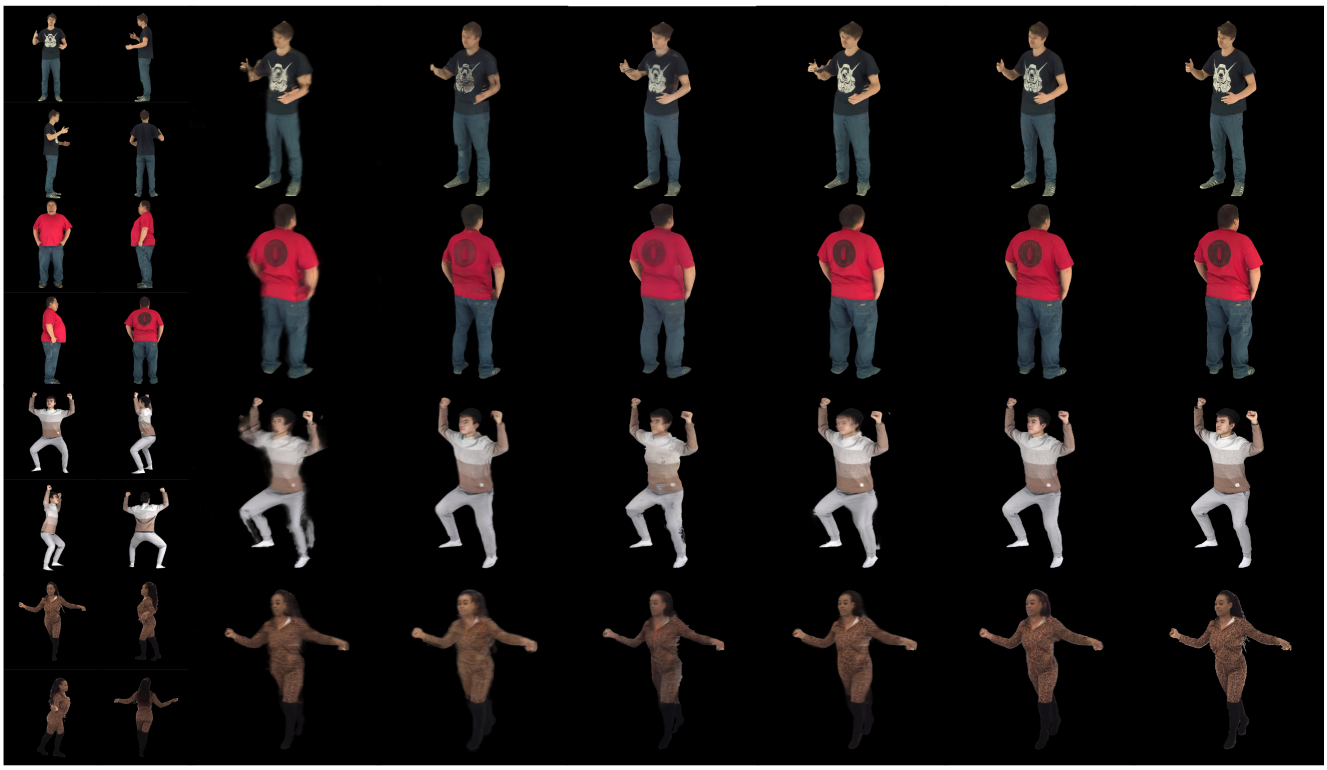

Figure 3. **Qualitative comparison with generalizable NeRFs**. We input $m = 4$ multi-view images of unseen identity, and our method produces a more photo-realistic novel view image compared to other state-of-the-art generalizable human NeRFs [8, 19, 26, 47]. The first two rows are from Multi-Garment [4], the third row from THuman2.0 [55] and the last row from GeneBody [8].

	Multi-Garment [4]			THuman2.0 [55]			ZJUMocap [31]			GeneBody [8]		
Model	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet [47]	28.44	0.924	0.0917	25.66	0.916	0.1033	25.25	0.876	0.2323	**24.71**	0.889	0.1364
NHP [19]	26.04	0.925	0.0701	26.99	0.935	0.0734	25.92	0.904	0.1623	22.75	0.872	0.1659
GNR [8]	28.61	0.937	0.0511	25.82	0.929	0.0605	25.39	0.903	0.1306	22.21	0.887	0.1254
KeypointNeRF [26]	28.36	0.938	0.0471	25.93	0.929	0.0607	25.85	0.910	0.1092	24.34	0.902	0.1236
Ours	**30.18**	**0.947**	**0.0305**	**28.88**	**0.952**	**0.0335**	**26.74**	**0.919**	**0.0955**	23.90	**0.906**	**0.0865**

Table 1. **Quantitative comparisons with the generalizable NeRF methods**. We evaluate the novel view synthesis performance on the unseen identity of different datasets. Our method significantly outperforms the state-of-the-art methods.

Normal Regularization. Although NeRF [27] can produce realistic images, the geometric surfaces generated by Marching Cubes [24] are extremely coarse and noisy. To alleviate it, we introduce normal regularization to constrain the normal among adjacent points.

$$\begin{aligned} \mathcal{L}_n &= \sum_{\mathbf{x}_s \in \mathcal{S}} \left\| \mathbf{n}\left(\mathbf{x}_s\right) - \mathbf{n}\left(\mathbf{x}_s + \epsilon\right) \right\|_2 \\ \mathbf{n}\left(\mathbf{x}_s\right) &= \frac{\nabla_{\mathbf{x}_s} \sigma\left(\mathbf{x}_s\right)}{\left\| \nabla_{\mathbf{x}_s} \sigma\left(\mathbf{x}_s\right) \right\|_2} \end{aligned} \tag{11}$$

where $\mathcal{S}$ is the points set randomly sampled near the SMPL mesh surface. $\mathbf{n}\left(\mathbf{x}_s\right)$ is the normal of the sampled point $\mathbf{x}_s$ and ϵ is a gaussian random noise with a variance of 0.1.

The final loss can be summarized as

$$\mathcal{L}_{full} = \mathcal{L} + \lambda_p \mathcal{L}_p + \lambda_n \mathcal{L}_n \tag{12}$$

where λ_p and λ_n are the weights of the perceptual loss and the normal regularization, respectively.

3.5. Efficient Training

During training, we select m multi-view images $\{I_k\}_{k=1}^m$ as inputs to build the generalizable model-based neural radiance fields and synthesize the target image I_t with given camera pose. It is memory-consuming to synthesize the whole image at the same time by volume rendering, so we only generate an image patch of the resolution $H_p \times W_p$ sampled randomly from the whole target image, which

Figure 4. **Qualitative results of novel pose synthesis on ZJUMocap [31] datasets**. Fs denotes training from scratch, Ft indicates fine-tuning the model after pretraining on Multi-Garment [4] dataset.

	Novel View Synthesis			Novel Pose Synthesis		
Model	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NB [31]	28.30	0.9462	0.0951	23.86	0.8971	0.1427
Ani-N [30]	26.19	0.9213	0.1399	23.38	0.8917	0.1594
A-NeRF [41]	27.43	0.9379	0.1019	22.40	0.8629	0.1991
ARAH [48]	**28.51**	**0.9483**	0.0813	24.63	0.9112	0.1070
Ours (Fs)	27.56	0.9314	0.0904	26.68	0.9241	0.0984
Ours (Ft)	28.45	0.9419	**0.0733**	**27.63**	**0.9361**	**0.0798**

Table 2. **Quantitative comparisons with case-specific optimization methods on ZJUMocap dataset.**

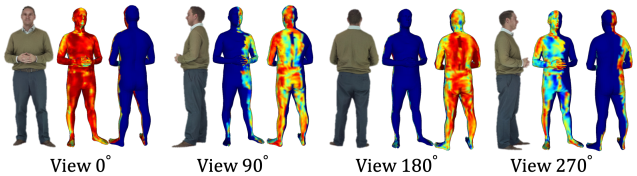

Figure 5. **The visualization of visibility-based attention confidence**. We visualize the contribution of different input views to the SMPL vertices. (Red indicates high confidence, while blue represents low confidence.)

means we only need to synthesize the half intermediate feature image of the resolution $\frac{H_p}{2} \times \frac{W_p}{2}$ by volume rendering. Meanwhile, since the perceptual loss requires a large enough image as input, we use partial gradient backpropagation introduced in CIPS-3D [58] to further reduce the memory cost caused by volume rendering. Specifically, we randomly choose n_p camera rays to participate in the gradient calculation, and the remaining rays $\frac{H_p}{2} \times \frac{W_p}{2} - n_p$ are not involved in gradient backpropagation.

4. Experiments

4.1. Datasets

We conduct experiments on two synthesized datasets Thuman2.0 [55] and Multi-garment [4] and real-world datasets Genebody [8] and ZJUMocap [31] for the generalizable scene task. The Thuman2.0 dataset contains 525 human scan meshes, of which we selected 400 for training and the remaining 125 for testing. For the Multi-garment dataset, we used 70 meshes for training and 25 meshes for evaluation. For each scanned mesh, we rendered it into 66 multi-view images of resolution 1024×1024. Specifically, we first place each scanned mesh into the center of a unit sphere at a distance of $5.4m$, with the camera orientation always pointing towards the center of the sphere. We move the camera around the sphere, sample a yaw angle from $0°$ to $60°$ with an interval of $30°$, and sample a roll angle from $0°$ to $360°$ with an interval of $30°$. The Genebody consists of 50 sequences at a 48 synchronized cameras setting, each of which has 150 frames. Specifically, we choose 40 sequences for training and another 10 sequences for testing. For ZJUMocap, which captures 10 dynamic human sequences with 21 synchronized cameras, we use 7 sequences for training and the rest 3 sequences for testing. To compare with the case-specific methods, we conduct experiments about novel view synthesis and novel pose synthesis on ZJUMocap. Following the evaluation protocols used in NB [31], we select 4 fixed view videos for training.

4.2. Evaluation Metrics

We evaluate our method with state-of-the-art generalizable or per-scene optimized methods to verify the superiority of our performance. We formulate comparative experiments on both geometric reconstruction and novel view synthesis. For quantitative comparison, we adopt peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS [56]) to evaluate the similarity between the rendered image and the ground-truth. Meanwhile, we also adopt chamfer distance (Chamfer) and point-to-surface distance (P2S) for geometric quality evaluation.

Figure 6. **Qualitative results of ablation studies on Multi-Garment dataset**.

Model	PSNR↑	SSIM↑	LPIPS↓
Base	25.22	0.895	0.1048
Base+MAB	27.08	0.915	0.0611
Base+PL	27.75	0.913	0.0673
Base+MAB+PL	28.72	0.929	0.0423
Base+MAB+NR	30.03	0.940	0.0562
Full Model	**30.18**	**0.947**	**0.0305**

Table 3. **Quantitative results of ablation studies on the Multi-garment**. Impact of the different components in our method.

4.3. Implementation Details

In our experiments, we choose $m = 4$ multi-view images $\{I_k \in \mathbb{R}^{512\times512\times3}\}_{k=1}^{m}$ as input to synthesize the target image $I_t \in \mathbb{R}^{512\times512\times3}$. During training, the input multi-view images are selected randomly, while selected uniformly surrounding the person (*i.e.*, the front, back, left, and right views) for evaluation. The resolution of the patch image during training is $H_p = W_p = 224$. The SMPL parameters are obtained using EasyMocap [31]. The size of our 3D feature volume $\mathcal{G}$ is 224^3. For partial gradient backpropagation, we randomly sample $n_p = 4,096$ camera rays from the target image patch to improve memory efficiency. We then uniformly query $N = 64$ samples from our feature volume along the camera ray. We train our network end-to-end by using the Adam [17] optimizer, and the base learning rate is 5×10^{-4} which decays exponentially along with the optimization. We train $200,000$ iterations on two Nvidia RTX3090 GPUs with a batch size of 4. The loss weights $\lambda_r = 1$, $\lambda_s = 0.1$, $\lambda_p = 0.01$, $\lambda_n = 0.01$.

4.4. Evaluation.

Comparison with generalizable NeRFs. We compare our method with state-of-the-art generalizable methods IBRNet [47], NHP [19], GNR [8] and KeypointNeRF [26]. We retrain all aforementioned networks with the official training protocols on GeneBody [8], Multi-Garment [4], and THuman2.0 [55] datasets. Specially, we also use $m = 3$ views as input on ZJUMocap [31] dataset following the evaluation protocol used in KeypointNeRF. The result can be seen in Tab. 1 and Fig. 3, which shows our method generalizes to unseen identities well and outperforms the methods compared. IBRNet, which learns a general view interpolation function to synthesize the novel view from a sparse set of nearby views, is able to render high-fidelity images for views close to the input views while having very poor generalization for views far from the input views. Our method has better generalization of novel view synthesis and generates higher quality geometry due to the use of the geometry prior SMPL. KeypointNeRF utilizes sparse 3D keypoints as pose priors and has weak expressiveness for unseen poses when the pose diversity in the training set is insufficient. In our experiment, we choose the 3D joints of SMPL as the input of KeypointNeRF. Compared to NHP and GNR, although we both employ SMPL as the geometry prior and suffer from inaccurate SMPL estimation, our method can alleviate the ambiguity of misalignment between geometry and pixel-aligned appearance. Meanwhile, benefiting from

Figure 7. **Qualitative results of different multi-view fusion mechanisms**.

Model	PSNR↑	SSIM↑	LPIPS↓
Ours with AvgPool	27.81	0.9317	0.05059
Ours with IBRAtt	28.50	0.9345	0.03413
Ours	**28.88**	**0.9518**	**0.03349**

Table 4. **Quantitative results of different multi-view fusion mechanisms on the THuman2.0 dataset**. AvgPool is used in PIFu [33] and PixelNeRF [54]. IBRAtt is proposed by IBRNet.

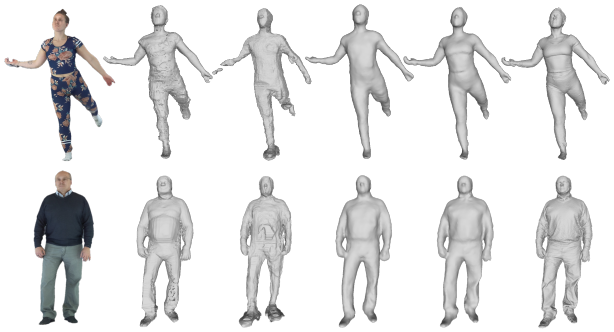

Figure 8. **Visualization results of 3D geometry reconstruction compared with different methods**.

	Multi-Garment [4]		THuman2.0 [55]	
Model	Chamfer↓	P2S↓	Chamfer↓	P2S↓
GNR [8]	1.3570	1.8981	1.7899	2.5932
NHP [19]	1.4646	2.2438	1.6027	2.3921
Ours	0.7175	0.6919	0.7444	0.6600
Ours(Ft)	**0.3721**	**0.3676**	**0.5172**	**0.4506**

Table 5. **Quantitative comparisons of 3D geometry reconstruction**. Our method consistently outperforms other methods, capturing more local details after fine-tuning.

perceptual loss, our generated images have more photorealistic local details. For 3d reconstruction, the mesh surface extracted by Marching Cubes is smoother and more precise due to normal regularization compared with others as shown in Fig. 8 and Tab. 5.

Comparison with case-specific Methods. We also compare with per-scene optimization methods NB [31], Ani-NeRF [30], A-NeRF [41], ARAH [48]. NB optimizes a set of structured latent codes associated with SMPL vertices, which are diffused into the observation space by using SparseConvNet. Since the 3D convolution in SparseConvNet is not rotation-invariant, NB has poor generalization on out-of-distribution poses. Ani-NeRF learns a backward LBS network to warp the observation space into the canonical space, which is not sufficient to model non-rigid deformations in complex poses. A-NeRF uses skeleton-relative embedding to model pose dependence deformation, which requires seeing the subjects from all views in varying poses. ARAH uses iterative root-finding for simultaneous ray-surface intersection search and correspondence search, which generalizes well to unseen poses. As shown in Tab 2, the performance of novel view synthesis is comparable with these methods, and it is reasonable since our network has more parameters(13.6M) and struggles with overfitting when the training data is so limited without any pretraining. After pretraining on the Multi-Garment and finetuning $5,000$ steps on ZJUMocap, our results achieve a noticeable improvement. Anyway, our method has superior generalization on novel pose synthesis, which is a more challenging task. Our results are more photorealistic and preserve more details like wrinkles and patterns as shown in Fig 4, which benefit from the sparse multi-view input.

4.5. Ablation studies

The baseline (**Base**) is an extended version of NB to express arbitrary identities as our baseline. Specifically, our structured latent codes are obtained by fusing multi-view input, rather than optimizing from scratch for a specific identity. Beyond that, we introduce multi-view appearance blending (**MAB**), perception loss (**PL**), and neural rendering (**NR**). The experimental results prove the effectiveness of each component as shown in Tab. 3 and Fig. 6. In addition, we explore the effects of different multi-view fusion mechanisms, and the experiments prove that our proposed visibility-based attention and geometry-guided attention are more effective than AvgPool [33, 54] and IBRAtt [47].

5. Limitations

There are some limitations of our method that need to be improved: i) Due to the minimal-clothed topology of SMPL, our model struggles to express extremely loose clothes and accessories. ii) When the testing pose is out-of-distribution, our method may produce some artifacts in results since 3D convolution in SparseConvNet [13] is not rotation-invariant. iii) As the target view moves further away from the input views, artifacts tend to emerge in the unobserved areas.

6. Conclusion

In this paper, we propose an effective framework to build generalizable model-based neural radiance fields (GM-NeRF) from sparse calibrated multi-view images of arbitrary performers. To improve generalization on novel poses and identities, we introduce SMPL as the structured geometric body embedding. However, inaccurate estimations of SMPL have a negative impact on the reconstruction results. To address this, we propose a novel geometry-guided multi-view attention mechanism that can effectively alleviate the misalignment between SMPL geometric prior feature and pixel-aligned feature. Meanwhile, we propose strategies such as neural rendering and partial gradient backpropagation to efficiently train our network using perceptual loss. Extensive experiments show that our method outperforms concurrent works.

Acknowledgements. The paper is supported in part by the National Key R&D Program of China under Grant No. 2018AAA0102001 and the National Natural Science Foundation of China under grant No.62293542, U1903215, and the Fundamental Research Funds for the Central Universities No.DUT22ZD210.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor S. Lempitsky. Neural point-based graphics. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12367 of *Lecture Notes in Computer Science*, pages 696–712. Springer, 2020. 2

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus A. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, pages 2293–2303. IEEE, 2019. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005. 2

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, pages 5419–5429, 2019. 2, 5, 6, 7, 8

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021. 1, 2

[6] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular RGB videos. *CoRR*, abs/2106.13629, 2021. 2

[7] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, volume 13683 of *Lecture Notes in Computer Science*, pages 222–239. Springer, 2022. 1, 2

[8] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *CoRR*, abs/2204.11798, 2022. 1, 2, 4, 5, 6, 7, 8

[9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, volume 12355 of *Lecture Notes in Computer Science*, pages 20–40. Springer, 2020. 2

[10] Paul E. Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In Judith R. Brown and Kurt Akeley, editors, *SIGGRAPH*, pages 145–156. ACM, 2000. 1

[11] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *PAMI*, 2021. 2

[12] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip L. Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts-Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. Fusion4d: real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):114:1–114:13, 2016. 1

[13] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232. Computer Vision Foundation / IEEE Computer Society, 2018. 4, 8

[14] Kaiwen Guo, Peter Lincoln, Philip L. Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6):217:1–217:19, 2019. 1

[15] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *CVPR*, pages 3090–3099. Computer Vision Foundation / IEEE, 2020. 2

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 694–711. Springer, 2016. 4

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 7

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *CVPR*, pages 5252–5262. Computer Vision Foundation / IEEE, 2020. 2

[19] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NIPS*, 34, 2021. 1, 2, 4, 5, 7, 8

[20] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *CVPR*, pages 1341–1350. Computer Vision Foundation / IEEE, 2020. 1

[21] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. Vis. Comput. Graph.*, 16(3):407–418, 2010. 1

[22] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, 2019. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 1, 2, 3

[24] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In Maureen C. Stone, editor, *SIGGRAPH*, pages 163–169. ACM, 1987. 5

[25] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470. Computer Vision Foundation / IEEE, 2019. 2

[26] Marko Mihajlovic, Aayush Bansal, Michael Zollhöfer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. *CoRR*, abs/2205.04992, 2022. 1, 2, 5, 7

[27] 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. 2, 4, 5

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174. Computer Vision Foundation / IEEE, 2019. 2

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 2

[30] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14294–14303. IEEE, 2021. 2, 6, 8

[31] 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*, pages 9054–9063, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[32] Sergey Prokudin, Michael J. Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *WACV*, pages 1809–1818. IEEE, 2021. 1

[33] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314. IEEE, 2019. 2, 7, 8

[34] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 81–90. Computer Vision Foundation / IEEE, 2020. 2

[35] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, pages 15851–15861. IEEE, 2022. 2

[36] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor S. Lempitsky. Textured neural avatars. In *CVPR*, pages 2387–2397. Computer Vision Foundation / IEEE, 2019. 1

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 4

[38] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020. 2

[39] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, pages 2437–2446. Computer Vision Foundation / IEEE, 2019. 2

[40] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS*, pages 1119–1130, 2019. 2

[41] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2, 6, 8

[42] 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. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020. 2

[43] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1, 2

[44] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 209–217. IEEE Computer Society, 2017. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 5998–6008, 2017. 4

[46] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul E. Debevec, Jovan Popovic, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.*, 28(5):174, 2009. 1

[47] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4690–4699. Computer Vision Foundation / IEEE, 2021. 1, 2, 5, 7, 8

[48] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 1–19. Springer, 2022. 2, 6, 8

[49] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Nikolai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-

Bianchi, and Roman Garnett, editors, *NeurIPS*, pages 1152–1164, 2018. 1

[50] Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus H. Gross. Scalable 3d video of dynamic scenes. *Vis. Comput.*, 21(8-10):629–638, 2005. 1

[51] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16189–16199. IEEE, 2022. 2

[52] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *CVPR*, pages 1679–1688, 2020. 1, 2

[53] Guangming Yao, Hongzhi Wu, Yi Yuan, Lincheng Li, Kun Zhou, and Xin Yu. Learning implicit body representations from double diffusion based neural radiance fields. In Luc De Raedt, editor, *IJCAI*, pages 1566–1572. ijcai.org, 2022. 2

[54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 1, 2, 4, 7, 8

[55] 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*, pages 5746–5756, 2021. 5, 6, 7, 8

[56] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 6

[57] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *PAMI*, 44(6):3170–3184, 2022. 2

[58] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *CoRR*, abs/2110.09788, 2021. 6