

GHuNeRF: Generalizable Human NeRF from a Monocular Video

Chen Li Jihao Lin Gim Hee Lee

Department of Computer Science, National University of Singapore

lichen@u.nus.edu

jihao.lin@u.nus.edu

gimhee.lee@comp.nus.edu.sg

Abstract

In this paper, we tackle the challenging task of learning a generalizable human NeRF model from a monocular video. Although existing generalizable human NeRFs have achieved impressive results, they require multi-view images or videos which might not be always available. On the other hand, some works on free-viewpoint rendering of human from monocular videos cannot be generalized to unseen identities. In view of these limitations, we propose GHuNeRF to learn a generalizable human NeRF model from a monocular video of the human performer. We first introduce a visibility-aware aggregation scheme to compute vertex-wise features, which is used to construct a 3D feature volume. The feature volume can only represent the overall geometry of the human performer with insufficient accuracy due to the limited resolution. To solve this, we further enhance the volume feature with temporally aligned point-wise features using an attention mechanism. Finally, the enhanced feature is used for predicting density and color for each sampled point. A surface-guided sampling strategy is also introduced to improve the efficiency for both training and inference. We validate our approach on the widely-used ZJU-MoCap dataset, where we achieve comparable performance with existing multi-view video based approaches. We also test on the monocular People-Snapshot dataset and achieve better performance than existing works when only monocular video is used.

1. Introduction

Free-viewpoint synthesis of human performers has wide applications such as virtual reality, movie production, gaming, etc. Traditional methods generally rely on images captured from dense camera views [7, 10] or accurate depth information [6, 8, 34], which are tedious and expensive to obtain. Recently, neural radiance fields (NeRF) [26] is proposed to represent a scene as a continuous 5D function, and has achieved high-fidelity rendering results. However, vanilla NeRF inherits the limitation of the requirement for dense camera views, and furthermore requires computation-

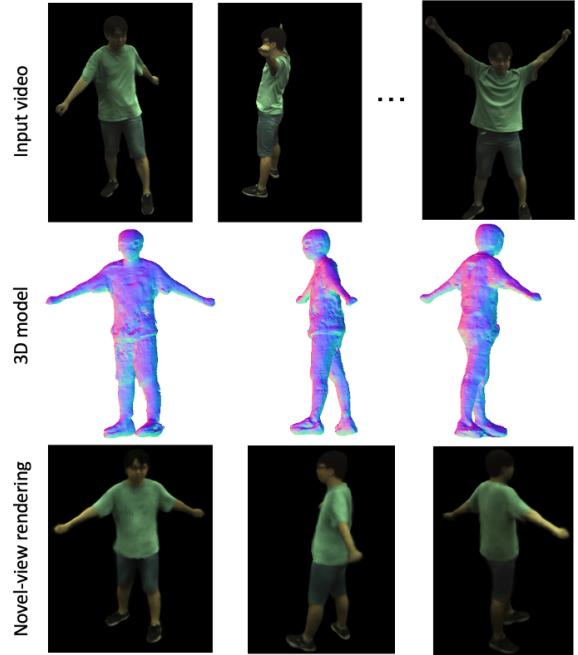


Figure 1. An illustration of our task. We aim to construct a 3D human NeRF model that can be used to render free-viewpoint images from a monocular video of a performer.

ally costly and non-generalizable per-scene optimization.

Recent human NeRFs [29, 38, 14] tackle the first limitation by aggregating information from videos to compensate for the lack of dense-view images. The aggregation of information is achieved by attaching a set of latent codes to the SMPL vertices [25] or by constructing a canonical space. Impressive results have been achieved by these methods with a very sparse-view setting or even monocular videos. However, these methods still need per-subject training and cannot generalize to unseen subjects during test. To mitigate the generalization issue, generalizable human NeRFs [42, 20, 5] have been proposed to avoid per-subject training. The core idea is to take pixel-aligned features as input instead of the position information in the original NeRF formulation. Although sparse-view synthesis

has been achieved by leveraging prior knowledge from pre-learned SMPL model, existing generalizable human NeRFs still require multi-view images or videos for both training and test. Unfortunately, this multi-view set-up is not always available in practice. In this paper, we aim to learn a generalizable human NeRF from monocular videos to overcome both the generalization issue and multi-view limitation. An illustration of our task is shown in Figure 1.

One challenge of learning human NeRF from monocular videos is the modeling of large human motions. To this end, we make use of the parametric SMPL model to construct a feature volume. The feature representation for each SMPL vertex can be obtained by projecting it into 2D image space. Given that a body part is not always visible across the whole video due to occlusion, we propose a visibility-aware feature aggregation to extract useful information from each observed frame. This vertex-wise feature is then diffused to the whole feature volume by applying SparseConvNet [9]. However, the feature volume can only represent the overall human geometry with insufficient accuracy because of the sparsity of the SMPL vertices and the limited volume resolution. To overcome this, we further enhance the volume feature at each location with a point-wise image feature. The point-wise feature is easy to obtain in the *multi-view setting* by projecting a 3D point in the target space to the observed views. In contrast, we cannot directly use the 3D-to-2D projection in our *monocular video setting* since the human is moving across the video and the corresponding 3D point in the observed space is unknown. We solve this issue by learning a transformation mapping from the target frame to the observed frames. The transformation is computed based on Linear Blend Skinning [22], where the blend weights are initialized with the SMPL model and further refined with a refinement network. Finally, we fuse the volume feature and the point-wise feature with an attention mechanism.

We further propose a surface-guided sampling strategy to improve the efficiency for both training and inference. Instead of randomly sampling points along a ray between a near and a far point as done in the vanilla NeRF, we sample points around the surface region to save memory and computation. Moreover, this also helps to regularize the 3D geometry implicitly since we are assuming that far away regions are empty space. We demonstrate the effectiveness of our approach on the widely used ZJU-MoCap dataset [29], where we achieve comparable performance with existing multi-view video based approaches. We also test on the monocular People-Snapshot dataset [1] and achieve better performance. Our main contributions are as follows:

- To the best of our knowledge, we are the first to tackle the task of learning a generalizable human NeRF model from monocular videos.
- We introduce GHuNeRF which consists of a visibility-

aware volume feature aggregation and temporal aligned feature enhancement to aggregate information across video frames for free-viewpoint image synthesis.

- We achieve state-of-the-art performance when only monocular video is available, and comparable performance with existing approaches that use multi-view videos.

2. Related Work

3D human reconstruction. 3D human reconstruction from images is an extensively studied problem in computer vision. Early works [3, 21, 19, 18, 27] reconstruct 3D human body by fitting a parametric 3D human models [25, 16] to the input data. The parametric model makes it possible to recover both shape and pose information with only 2D supervision such as keypoints, silhouette and body segmentations. However, The parametric models are learned from minimally clothes body data, and hence cannot generalize well to clothed people. To solve this, implicit representation based approaches [31, 32, 13, 12] are proposed which can represent various topologies. PIFu and its variants [31, 32] adopt the occupancy field and reconstruct detailed surface geometry from even one view. ARCH and ARCH++ [13, 12] combine the parametric model with the implicit field to estimate animatable 3D human avatars. Despite the high-fidelity reconstruction, these methods require the 3D ground truth for training which are expensive to collect. In comparison, we aim to achieve 3D human reconstruction and rendering from monocular videos.

Neural scene representation and rendering. Recently, various neural representations [24, 26, 35, 43, 39] have been introduced for novel view synthesis and geometric reconstruction. In particular, Neural radiance field (NeRF) [26] represents a scene as a continuous 5D function and has achieved high-fidelity rendering results. Although effective, one main limitation of the vanilla NeRF is the requirement for expensive per-scene optimization. To mitigate this, subsequent works [4, 15, 37, 41] learn generalizable NeRFs across different scenes by taking the pixel-aligned features as conditional information. Specifically, MVSNeRF [4] leverages plane-swept cost volumes to construct a neural encoding volume with per-voxel neural feature. An MLP is then adopted to regress volume density and color by using features interpolated from the encoding volume. IBRNet [37] introduces a ray transformer to aggregate information from nearby source views along a given ray. Inspired by these works, we also design a generalizable NeRF to model human body across different identities. This is more challenging comparing to existing generalizable NeRF [4, 15, 37, 41] for static scenes since we need to model the large motion of human body simultaneously.

Neural radiance fields for human. To model the motion of human body, existing human NeRFs [33, 40, 29, 23, 14, 28, 38] rely on human-prior information, such as a skeleton or a parametric model. NeuralBody [29] attaches a set of latent codes to the vertices of the SMPL model, which is able to aggregate information across different video frames. The per-vertex latent code is then diffused to generate a continuous latent code volume, which is used to regress the density and color for volume rendering. The other line of works [40, 23, 14, 28, 38] map all observations to a shared canonical space to model the large deformation of human bodies. HumanNeRF [38] adopts the inverse linear-blend skinning, which is combined with a non-rigid deformation, to learn a motion field mapping from observation to canonical space. Impressive results have been achieved by HumanNeRFs even when using the monocular video as the inputs. However, these methods are designed for the person-specific setup, where a model needs to be trained for each identity. To solve this problem, recent works [42, 20, 5] consider generalizable human rendering by taking the pixel-aligned features as conditions to avoid the memorization of human-specific density and color. GPNeRF [5] proposes a geometry-guided progressive rendering mechanism, which leverages the geometry volume and the predicted density to reduce the number of sampling points. The most related work is NHP [20], which designs a temporal transformer to aggregate tracked visual features based on the skeletal body motion. The temporally-fused features are then merged with multi-view pixel-aligned features for novel view synthesis. Although NHP achieves cross-identity and cross-dataset generalization, it requires multi-view videos which might not be always available in practice. In this paper, we aim to learn a generalizable human NeRF model with only monocular video as the inputs.

3. Our Method: GHuNeRF

We propose GHuNeRF to construct a 3D human model that can be used to render free-viewpoint images from a monocular video of a performer. The overall pipeline of our GHuNeRF is shown in Figure 2. To handle the large motions of human body, we leverage the SMPL model to aggregate vertex-wise feature from input video frames. Specifically, we introduce a **visibility-aware aggregation** scheme since a vertex can be observed from only some of the frames due to occlusion. We then construct a feature volume from the SMPL vertices features using SparseConvNet. The volume information can only represent the overall geometry, but is not sufficiently accurate due to the sparsity of the SMPL vertices and the limited resolution of the volume. To solve this, we further enhance the volume features with **temporally aligned** point-wise features based on an attention mechanism. This enhanced feature is used to predict the color and density for novel-view rendering.

Moreover, we also introduce a **surface-guided point sampling strategy**, where only points around the surface region are sampled to improve the efficiency.

3.1. Visibility-aware Volume Feature Aggregation

We denote the observed video as $I_{1:T} = \{I_1, I_2, \dots, I_T\}$, where T is the number of frames. The objective is to reconstruct a 3D human model that can be rendered at any time step t at any camera view c , which we denote as target frame I_g . To synthesize the target frame, we construct a feature volume in the space of the target frame, *i.e.* target space, based on the target SMPL parameters. Specifically, for each vertex on the target SMPL, we obtain the feature representation by aggregating information from the observed frames. Visibility information is incorporated during the aggregation since a vertex can be visible to only some of the observed frames due to occlusion. The aggregated feature for each vertex is computed as:

$$F(v_g) = \frac{\sum_{i=1}^T s_i \times F(v_i)}{\sum_{i=1}^T s_i}, \quad (1)$$

where v_g denotes a vertex on the target SMPL and $\{v_1, v_2, \dots, v_T\}$ denote the corresponding SMPL vertex of the observed frames, s_i represents the visibility of the vertex v_i . The feature representations of the vertices in the observed frames are computed by projecting them into the 2D image space, where the images features are obtained using a 2D feature extractor. The SMPL model for each frame is pre-computed with [16] and the visibility information for each vertex is obtained from a rasterization process. By considering the visibility, we are able to collect useful information across different times steps which thus compensates for the lack of multi-view information.

The vertex-wise feature representation is sparse and does not fulfill the requirement for volume rendering, where the density and color are queried in a continuous 3D space. We further use the SparseConvNet [9] to diffuse the vertex-wise feature to the nearby 3D space following [29]. Specifically, a 3D bounding box of the human in the target frame is computed based on the corresponding SMPL parameters, and the box is gridded into a tessellation of smaller voxels with voxel size of $5\text{mm} \times 5\text{mm} \times 5\text{mm}$. The feature representation of any non-empty voxel is computed as the mean of features of SMPL vertices within this voxel.

3.2. Temporally Aligned Feature Enhancement

The feature volume constructed based on SMPL vertices can represent the overall structure of a human, but is not sufficiently accurate due to the sparsity of the SMPL vertices and the limited volume resolution. To overcome this problem, we propose to enhance the volume features with temporally aligned point-wise features. In the multi-view

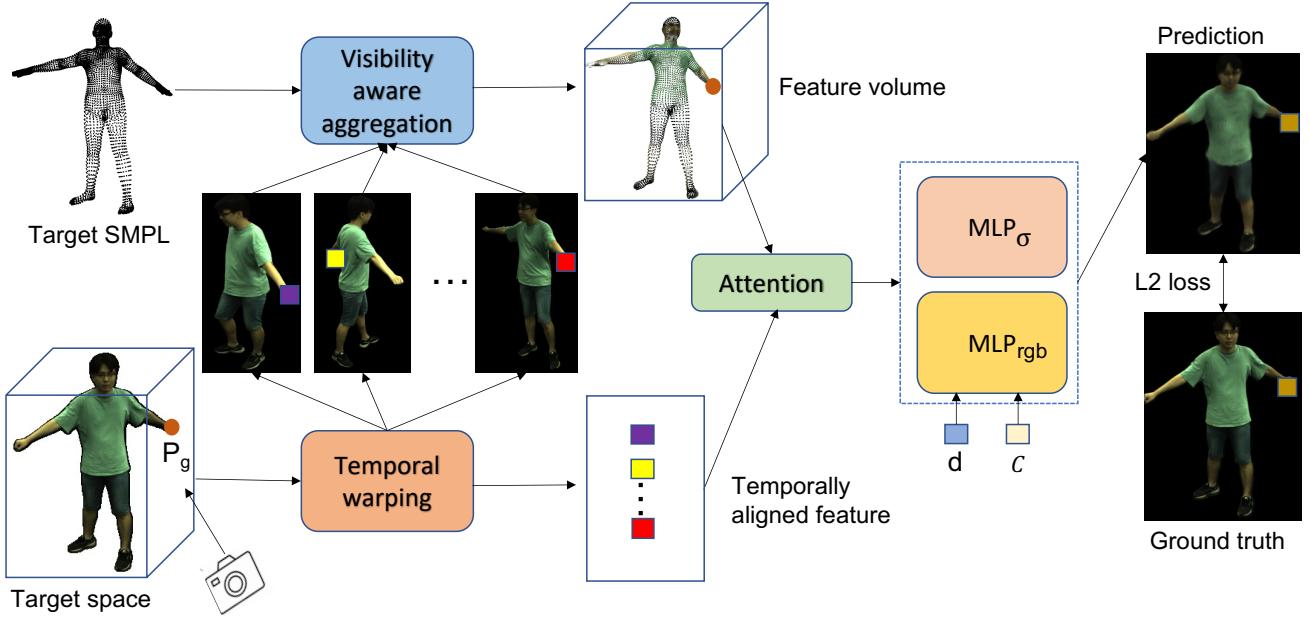


Figure 2. The overall pipeline of our GHuNeRF framework. We first compute the feature representation for each vertex of the target SMPL using the visibility-aware feature aggregation. A feature volume is then constructed based by diffusing the vertex-wise feature to nearby 3D space. The volume feature is further enhanced with the temporally aligned point-wise feature using the attention mechanism. Finally, the enhanced feature is used for predicting density and color for rendering.

based approaches [42, 20, 5], this point-wise feature can be easily obtained by projecting any sampled point in the target space into observed views with the known camera parameters. However, in our monocular video setting, we cannot directly project any 3D sampled point in the target space to the observed video frames since we do not know the corresponding 3D point in the observation space. To solve this, we define a transformation field mapping 3D sampled points from the target space to the observation space.

It is hard to directly model the transformation with deep networks given the large motion of human body. We instead formulate the transformation based on Linear Blend Skinning. The Linear Blend Skinning is commonly used in character animation, where the transformation of each vertex is affected by all body parts. Specifically, a vertex v on a template mesh can be posed via the Linear Blend Skinning as:

$$v' = \left(\sum_{j=1}^N w_j \mathcal{T}_j \right) v. \quad (2)$$

\mathcal{T}_j is the transformation for the j^{th} body part from a template pose and w_j represents the corresponding blend weight. N denotes the total number of body parts. In our case, the transformation \mathcal{T} for each body part can be computed from the pose parameter Θ and the joint locations J in the template pose based on SMPL. Given the target pose Θ_g and the observed pose Θ_o , the transformation from the target space

to the observation space can then be expressed as:

$$p_o = \left(\sum_{i=1}^N w_g \mathcal{T}_o \mathcal{T}_g^{-1} \right) p_g, \quad (3)$$

where w_g represents the blend skinning weights in target space. We follow previous works [28, 13, 2] to model the blending weights by leveraging prior knowledge from the SMPL model. Specifically, for any 3D sampled point in the target space, we assign the initial blending weight as the same value as the nearest vertex on the SMPL surface. These initial blending weights are generally inaccurate, especially for 3D points that are far from the surface. To improve the accuracy, we further use an MLP network for refinement, *i.e.*:

$$w_g = w_s + \text{MLP}_w(w_s, \Theta, d), \quad (4)$$

where w_s denotes the initial blending weight. MLP_w represents the refinement network, which takes the initial weight w_s , the pose parameters Θ and the distance to the nearest surface point d as inputs. Intuitively, the blend skinning weight of a 3D point in the target space depends on both the target body pose and the distance from this point to the body surface. We apply softmax to the output of the refinement network such that the weights for different body parts sum up to one.

For any 3D sampled point p_g in the target space, we can warp it to the observation spaces of all input frames

using Eqn. (3), denoted as $\{p_0^o, p_1^o, \dots, p_T^o\}$. The corresponding point-wise feature can then be extracted as $\{F(p_1^o), F(p_2^o), \dots, F(p_T^o)\}$ via a 3D-to-2D projection. We use these point-wise features to enhance the volume features. Specifically, the volume feature of a 3D point p_g can be retrieved from the volume using bilinear interpolation, which we denote as $F_v(p_g)$. This volume-based feature $F_v(p_g)$ is enhanced with the temporally aligned point-wise features with the attention mechanism [36]:

$$\begin{aligned} F_e(p_g) &= \text{Attention}(Q = F_v(p_g)), \\ K &= \{F(p_t^o)\}_{t=1}^T, \\ V &= \{F((p_t^o))\}_{t=1}^T. \end{aligned} \quad (5)$$

Intuitively, the attention mechanism helps to incorporate relevant information from the input frames and ignore the irrelevant ones.

Finally, the enhanced feature is used to predict the density and color of each sampled point:

$$\begin{aligned} \sigma(p_g) &= \text{MLP}_\sigma(F_e(p_g)), \\ c(p_g) &= \text{MLP}_{\text{rgb}}(F_e(p_g), \gamma_d(\mathbf{d}), \mathcal{C}), \end{aligned} \quad (6)$$

where MLP_σ and MLP_{rgb} are the density and color prediction networks, respectively, and γ_d represents the positional encoding for view direction d . \mathcal{C} is a per-camera latent code to encode the camera-specific elements.

3.3. Surface-guided Points Sampling

We further introduce a surface-guided sampling strategy to replace the random sampling used in the original NeRF. The motivation is twofold. First, the random sampling inevitably leads to numerous sampled points in the empty space and thus slowing down the convergence of the network. In contrast, our surface guided sampling strategy only samples points near the SMPL surface region, which significantly reduces the number of unnecessary points in the empty space and thus improving the efficiency. Second, our surface-guided sampling is able to regularize the 3D geometry implicitly since we are assuming that points far away from the surface corresponding to empty space. Computing the distance between a sampled point to the SMPL surface during training is expensive. To reduce computations, we voxelize the 3D space and pre-compute the distance between each voxel and the SMPL surface. The distance between any sampled 3D point and surface is computed with bilinear interpolation during training.

3.4. Volume Rendering

We use the volume rendering to render the RGB values for each pixel at time step t in the target view:

$$\tilde{C}_t(\mathbf{r}) = \sum_{k=1}^{N_k} T_k(1 - \exp(-\sigma_k \delta_k)) \mathbf{c}_k, \quad (7)$$

$$\text{where } T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j \delta_j\right). \quad (8)$$

N_k denotes the number of sampling points along each ray and δ_k is the distance between adjacent sampled points.

The objective function is the squared error between the rendered color $\tilde{C}_t(\mathbf{r})$ and the ground truth color $C_t(\mathbf{r})$:

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \|\tilde{C}_t(\mathbf{r}) - C_t(\mathbf{r})\|_2^2. \quad (9)$$

During training, our network can be supervised with both multi-view or monocular videos. In the multi-view training (MVT) setting, we aim to synthesis an image at any time step from a different view. In the monocular training (MoT) setting, we aim to synthesis an image at any time step in the same view. It should be noted that only a monocular video of a human performer is used as inputs during inference in both settings.

4. Experiments

4.1. Implementation details

Networks details. We use the ResNet18 [11] following [20] to extract image features, which are used to compute both vertex-wise feature in Eqn. (1) and point-wise feature in Eqn. (5). The blend weights refinement MLP consists of eight layers with the channel size of 256. The SparseConvNet consists of four blocks of convolution and down-sampling layers with $2\times$, $4\times$, $8\times$, $16\times$ downsampled sizes. The attention is performed twice for the density feature and color feature respectively. The per-camera latent code has a dimension of 128 and is optimized together with the network. Note that each video in both ZJU-MoCap and People-Snapshot datasets contains hundreds of frames, and we select 15 frames from the whole video as the inputs for both memory and computation efficiency. More details are provided in the supplementary material.

Training details. We adopt the Adam optimizer [17] for training with a learning rate of $1e^{-4}$ and a batch size of one. Both training and inference are conducted with an image size of 512×512 for fair comparison with previous works. We train our network for 500 epochs with 500 iterations in each epoch on one RTX 3090Ti GPU. The distance threshold for our surface-guided points sampling is set to 5 cm.

4.2. Datasets and Evaluation Metrics

We evaluate our approach on the ZJU-MoCap dataset [29] and the People-Snapshot dataset [1]. The ZJU-MoCap dataset consists of 9 dynamic human video with different appearance and poses. Each subject is captured using a multi-camera system with 21 synchronized cameras. We use 6 subjects for training and 3 subjects for evaluation. The People-Snapshot dataset is a monocular video dataset which captures performers that rotate while holding an A-pose. We randomly select 3 identities for testing and use the remaining for training. We evaluate our method under both MVT and MoT settings. We select 15 frames from the whole video sequence for each target image in the MVT setting, and always use the same 15 frames to synthesize all other frames in the same video in the MoT setting. We adopt the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as the evaluation metrics following [29, 20]. We also provide qualitative results for 3D human reconstruction as there is no ground truth human geometry.

4.3. Results: ZJU-MoCap Dataset

We first evaluate our method on the most commonly used ZJU-MoCap dataset. We conduct experiments under both MVT and MoT settings, where the network is trained with multi-view and monocular data, respectively. Note that we **only require a monocular video** as the input during test under both settings. We compare with previous human NeRF works including NT [35], NHR [39], NV [24], NB [29], PVA [30], pixelNeRF [41], GPNeRF [5] and NHP [20]. Among the comparison methods, NT, NHR, NV and NB are not generalizable which require per-scene optimization, PVA, GPNeRF and NHP are generalizable, but use multi-view videos (NHP) or images (GPNeRF, PVA) as inputs for both training and test. The results for pixelNeRF are obtained under multi-view setting although the input for pixelNeRF can be both monocular or multi-view images.

We test our approach on both seen and unseen identities and the quantitative results are shown in Table 1 and Table 2, respectively. Note that we use the same seen and unseen subjects as NHP [20], while GPNeRF uses a different split. As can be seen from Table 1, our approach achieves comparable performance with NHP, especially for the SSIM score, although we only use a monocular video as input during test. Moreover, we also achieve comparable results with the optimization based approach Neuralbody [29]. For the results for the unseen identities in Table 2, we outperform the optimization based approaches, which are trained on the unseen subjects before inference. Comparable performance is also achieved compared with NHP. Results in both tables verify the effectiveness of our approach in aggregating information from monocular videos.

We show the qualitative results for both seen and unseen identities in Figure 3, where we compare with NHP. We can

see that our approach is able to synthesize as high-fidelity images as NHP although NHP uses multi-view videos as the input. Even better results are achieved in some cases, where our approach generates more details on the face (marked with red box). We also show the 3D human model predicted by our network in Figure 4. Our approach is able to predict realistic 3D human shapes in the case where NHP fails to predict the left arm of the performer. More qualitative results are provided in the supplementary material.

Method	Generalizable	Monocular		Seen subjects	
		Train	Test	PSNR (\uparrow)	SSIM (\uparrow)
NT [35]	\times	\times	*	23.86	0.896
NHR [39]	\times	\times	*	23.95	0.897
NB [29]	\times	\times	*	28.51	0.947
NHP [20]	\checkmark	\times	\times	28.73	0.936
GPNeRF [5]	\checkmark	\times	\times	28.91	0.944
Ours-MVT	\checkmark	\times	\checkmark	27.24	0.930
Ours-MoT	\checkmark	\checkmark	\checkmark	27.32	0.936

Table 1. Quantitative results on the seen subjects of the ZJU-MoCap dataset. * indicates method is trained on per-scene optimization and does not require image input during test.

Method	Generalizable	Monocular		Unseen subjects	
		Train	Test	PSNR (\uparrow)	SSIM (\uparrow)
NV [24]	\times	\times	*	20.84	0.827
NT [35]	\times	\times	*	21.92	0.873
NHR [39]	\times	\times	*	22.03	0.875
NB [29]	\times	\times	*	22.88	0.880
PVA [30]	\checkmark	\times	\times	23.15	0.866
pixelNeRF [41]	\checkmark	\times	\times	23.17	0.869
NHP [20]	\checkmark	\times	\times	24.75	0.906
GPNeRF [5]	\checkmark	\times	\times	25.96	0.921
Ours-MVT	\checkmark	\times	\checkmark	24.12	0.905
Ours-MoT	\checkmark	\checkmark	\checkmark	24.55	0.911

Table 2. Quantitative results on the unseen subjects of the ZJU-MoCap dataset. * indicates method is trained on per-scene optimization and does not require image input during test.

4.4. Results: People-Snapshot Dataset

We also test our approach on the People-Snapshot dataset which only consists of monocular videos. Our approach is trained under the monocular setting (MoT) on this dataset since multi-view supervision is not available. We randomly select three subjects from the dataset as test subjects and the remaining are used for training. During inference, the same 15 frames are evenly selected from the whole video to synthesize the remaining video frames. We compare our approach with NHP [20] since NHP is also generalizable and can be trained with monocular videos. Note that the results of NHP in Table 3 are based on our implementation since the temporal aggregation component



Figure 3. Qualitative results for novel-view synthesis on the ZJU-MoCap dataset.

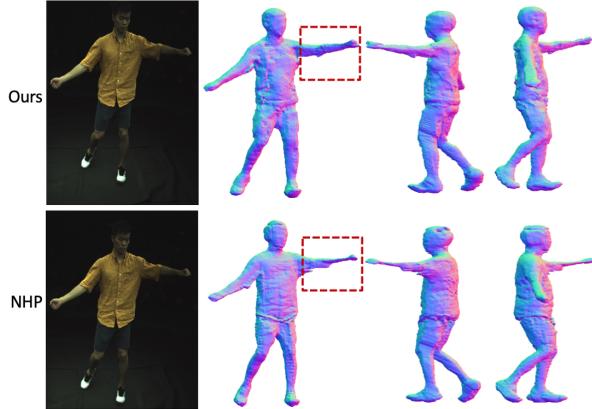


Figure 4. Qualitative results for 3D human model on the ZJU-MoCap dataset.

of NHP is not publicly available. As can be seen from Table 3, our approach outperforms NHP by a large margin when only monocular videos are available.

To evaluate the generalization capacity of our approach, we also show results for the cross-dataset generalization in Table 4. The results are obtained by directly applying our

model trained on the People-Snapshot dataset to the ZJU-MoCap dataset. We can see that the cross-dataset generalization achieves similar PSNR score with our model trained on the ZJU-MoCap dataset, 23.20 vs 24.55.

Method	Generalizable	Monocular		Test subjects	
		Train	Test	PSNR (\uparrow)	SSIM (\uparrow)
NHP [20]	✓	✓	✓	26.22	0.903
Ours-MoT	✓	✓	✓	28.37	0.927

Table 3. Quantitative results on the unseen subjects of the People-Snapshot dataset.

We show the qualitative results for novel-view synthesis on the People-Snapshot dataset in Figure 5. NHP fails to generate the correct color (marked in red box) for the target image when trained only with monocular videos. In comparison, our approach is able to synthesize realistic novel-view images across different identities. Qualitative results for the 3D human model predicted by our network are also provided in Figure 6. We can see that NHP struggles to estimate the 3D shapes, where holes on the chest of the performer can be observed. In contrast, our method estimates more realistic 3D human shapes with fewer artifacts.

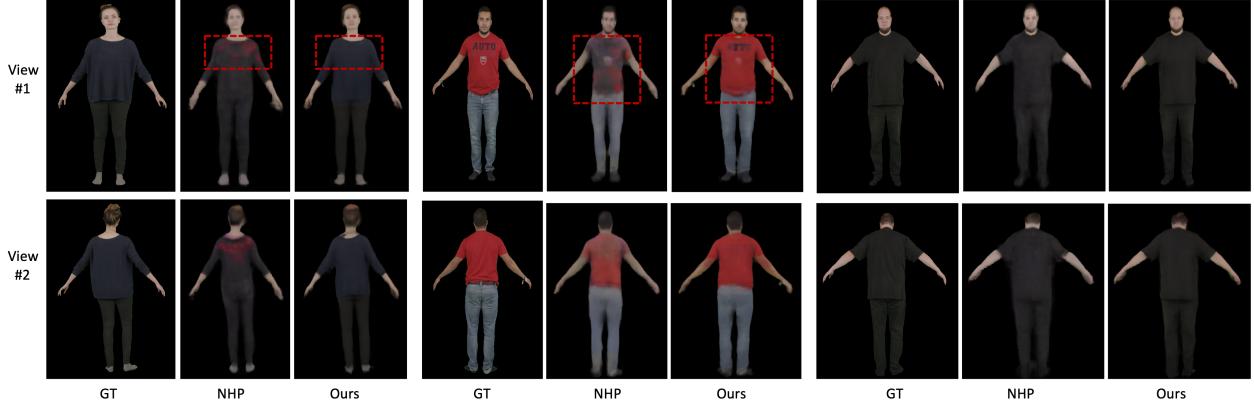


Figure 5. Qualitative results for novel view synthesis on the People-Snapshot dataset.

Method	Generalizable	Monocular Train Test	Test subjects	
			PSNR (\uparrow)	SSIM (\uparrow)
NHP [20]	✓	✓ ✓	16.07	0.836
Ours-MoT	✓	✓ ✓	23.20	0.889

Table 4. Quantitative results for cross-dataset generalization.

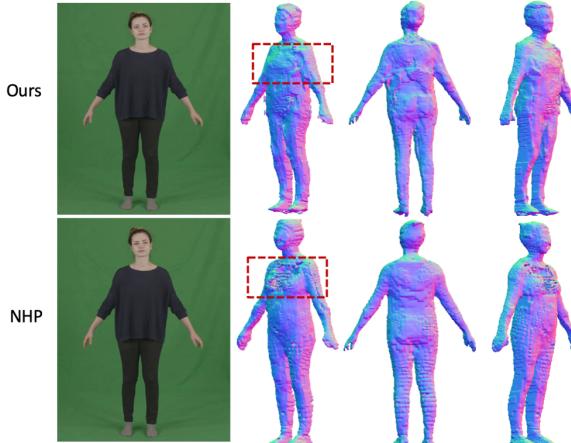


Figure 6. Qualitative results for 3D human model on the People-Snapshot dataset.

4.5. Ablation Studies

We conduct ablation studies on the ZJU-MoCap dataset by removing each component successively from the full model. The results are shown in Table 5, where ‘VVF’ denotes visibility-aware volume feature, ‘BWR’ denotes blending weights refinement, ‘TFE’ denotes temporally aligned feature enhancement and ‘SGS’ denotes surface-guided sampling. We first use the volume feature and remove the blending weights refinement, the temporally aligned feature enhancement and the surface-guided sampling successively. The performance drops when each component is removed, especially for the temporally aligned feature enhancement. To further verify the role of the

visibility-aware volume feature, we then remove the volume feature, and instead use the temporally aligned feature for density and volume prediction. We directly take the mean value of the temporal features instead of using the attention operation in Eqn.(5). We can see that the performance drops when the volume feature is removed.

	VVF	BWR	TFE	SGS	PSNR (\uparrow)	SSIM (\uparrow)
✓	✓	✓	✓	✓	24.12	0.905
✓	✗	✓	✓	✓	24.06	0.905
✓	✗	✗	✓	✓	22.94	0.896
✓	✗	✗	✗	✗	22.56	0.890
✗	✓	✓	✓	✓	23.70	0.888

Table 5. Ablation study of successive removal of each component.

5. Limitations

Our approach achieves impressive results on both novel-view synthesis and 3D human reconstruction, but there are still many challenges for the task of learning generalizable human NeRF from monocular videos. For example, the generalization capacity is still limited when the training and testing data are significantly different. Some failure cases are shown in our supplementary material.

6. Conclusion

We propose GHuNeRF to tackle the task of learning a generalizable human NeRF from monocular videos in this paper. We leverage the SMPL model to construct a feature volume where a visibility-aware feature aggregation scheme is introduced to integrate useful information from input frames. A volume feature enhancement is designed to enhance coarse volume feature with temporally-aligned point-wise feature. Moreover, a surface-guided sampling strategy is proposed to improve the efficiency for both training and inference. Extensive experiments have been conducted to validate the effectiveness of our approach.

Supplementary Material for GHuNeRF: Generalizable Human NeRF from a Monocular Video

Chen Li Jiahao Lin Gim Hee Lee

Department of Computer Science, National University of Singapore

lichen@u.nus.edu jiahao.lin@u.nus.edu gimhee.lee@comp.nus.edu.sg

Selection of video frames as the inputs. Each video in the ZJU-MoCap and People-Snapshot datasets contains hundreds of frames, which will cause memory issue if directly taking the whole video as input. To solve this problem, we select 15 frames from the whole video sequence based on two criteria: 1) To select evenly from each video, denoted as criterion #1. 2) To select based on the SMPL vertices, denoted as criterion #2. Specifically, we compute the distance between the SMPL vertices of the target frame and each video frame in the camera coordinate, and take the closest 15 frames as the inputs. Empirically, criterion #2 performs slightly better than criterion #1, as shown in Table 1. Intuitively, criterion #2 selects frames that have similar body direction and pose with the target frame, hence results in better performance.

Selection criteria	Seen subjects		Unseen subjects	
	PSNR (\uparrow)	SSIM (\uparrow)	PSNR (\uparrow)	SSIM (\uparrow)
Criterion #1	27.19	0.930	23.71	0.902
Criterion #2	27.24	0.930	24.12	0.905

Table 1. Quantitative results based on different selection criteria.

More qualitative results. We compare with NHP [20] for novel view synthesis on the ZJU-Mocap dataset. The results for both unseen and seen identities are shown in Figure 1 and 3, respectively. As marked in red box, we can see that our approach is able to generate more details on the face, arms and legs. We also show qualitative comparison for 3D reconstruction in Figure 4, where our method reconstructs more realistic shapes as marked in red box. We also provide more qualitative results for both novel view synthesis and 3D reconstruction of our method in the video.

Failure cases. Our proposed GHuNeRF achieves cross-dataset generalization as demonstrated in Section 4.4 of the main paper. However, the generalization capacity is still limited when the training and testing datasets are significantly different, as mentioned in the limitations. We show some examples in Figure 2, where we test our model trained on the People-Snapshot dataset on the ZJU-MoCap dataset. We can see that our method fails to predict the correct

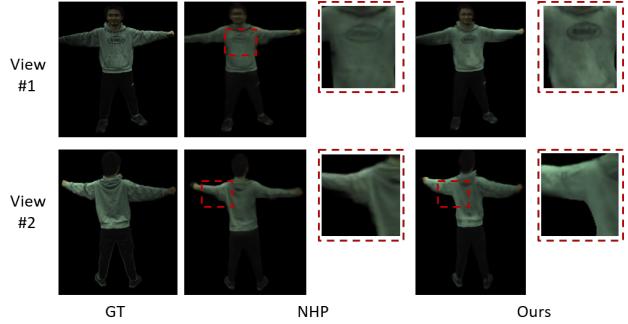


Figure 1. Qualitative comparison for unseen identities.

color on the face and legs. The main reason is the significant difference of lighting condition between the People-Snapshot dataset and the ZJU-MoCap dataset. The lighting for the ZJU-MoCap dataset is dark as can be seen from the ground truth images of the results for novel view synthesis, while the lighting for the People-Snapshot dataset is much brighter. Moreover, the yellow shirt has never been seen from the People-Snapshot dataset.

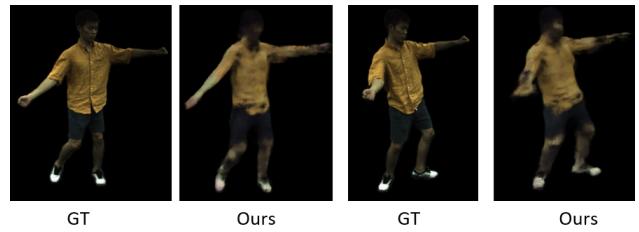


Figure 2. Examples of failure case.

References

- [1] Thimo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.
- [2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape

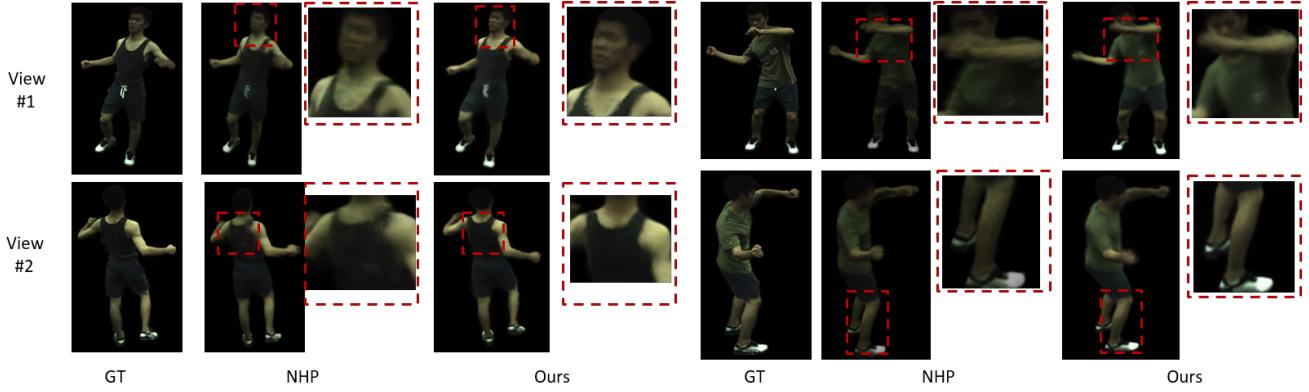


Figure 3. Qualitative comparison for seen identities.

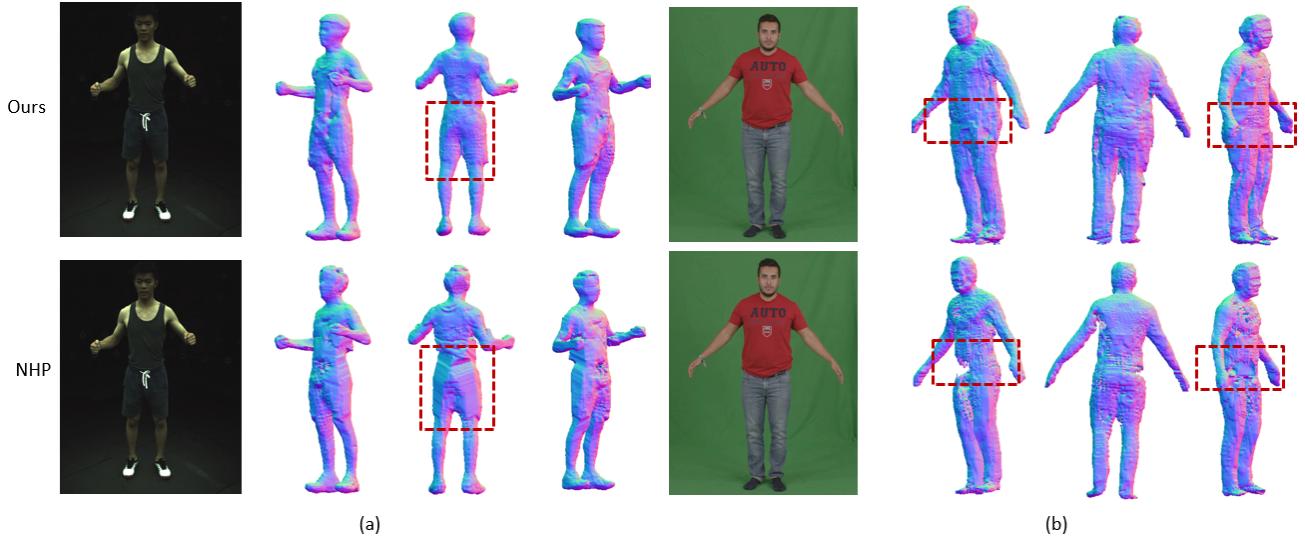


Figure 4. Qualitative comparison for 3D reconstruction on the ZJU-MoCap dataset (a) and People-Snapshot dataset (b).

- for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020.
- [3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016.
 - [4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.
 - [5] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pages 222–239. Springer, 2022.
 - [6] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.
 - [7] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000.
 - [8] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escalano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

- [9] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.
- [10] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escalano, Rohit Pandey, Jason Dourgarian, et al. The re-lightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)*, 38(6):1–19, 2019.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [12] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021.
- [13] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020.
- [14] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022.
- [15] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.
- [16] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018.
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019.
- [19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.
- [20] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021.
- [21] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017.
- [22] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000.
- [23] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021.
- [24] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.
- [25] 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.
- [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. *Communications of the ACM*, 65(1):99–106, 2021.
- [27] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.
- [28] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021.
- [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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.
- [30] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11742, 2021.
- [31] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigi, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.
- [32] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.

- [33] 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.
- [34] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 246–264. Springer, 2020.
- [35] 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.
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [37] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.
- [38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022.
- [39] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020.
- [40] Hongyi Xu, Thiendo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021.
- [41] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.
- [42] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7743–7753, 2022.
- [43] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.