DiHuR: Diffusion-Guided Generalizable Human Reconstruction

Jinnan Chen[1] Chen Li[2,3*] Gim Hee Lee[1]
Department of Computer Science, National University of Singapore[1]
IHPC, Agency for Science, Technology and Research, Singapore[2]
CFAR, Agency for Science, Technology and Research, Singapore[3]
jinnan.c@u.nus.edu lichen@u.nus.edu gimhee.lee@comp.nus.edu.sg

Abstract

We introduce DiHuR, a novel Diffusion-guided model for generalizable Human 3D Reconstruction and view synthesis from sparse, minimally overlapping images. While existing generalizable human radiance fields excel at novel view synthesis, they often struggle with comprehensive 3D reconstruction. Similarly, directly optimizing implicit Signed Distance Function (SDF) fields from sparse-view images typically yields poor results due to limited overlap. To enhance 3D reconstruction quality, we propose using learnable tokens associated with SMPL vertices to aggregate sparse view features and then to guide SDF prediction. These tokens learn a generalizable prior across different identities in training datasets, leveraging the consistent projection of SMPL vertices onto similar semantic areas across various human identities. This consistency enables effective knowledge transfer to unseen identities during inference. Recognizing SMPL's limitations in capturing clothing details, we incorporate a diffusion model as an additional prior to fill in missing information, particularly for complex clothing geometries. Our method integrates two key priors in a coherent manner: the prior from generalizable feed-forward models and the 2D diffusion prior, and it requires only multi-view image training, without 3D supervision. DiHuR demonstrates superior performance in both within-dataset and cross-dataset generalization settings, as validated on THuman, ZJU-MoCap, and HuMMan datasets compared to existing methods.

1. Introduction

Neural scene representation [16] enables the realistic generation of 3D digital models from 2D observations of the scene that are useful for many real-world applications such as Augmented and Virtual Reality (AR/VR), and the creation of digital avatars. Although existing generalizable human NeRF [4, 5, 8, 18, 23, 26] have achieved impressive, they mainly focus on novel view synthesis. The surfaces extracted from these fields are often unsatisfactory, particularly in sparse view settings. This is due to the little or non-existent overlap between views, which complicates feature aggregation. In such scenarios, it becomes challenging to identify the correct features with low variance(high confidence) across views. Consequently, the algorithm often resorts to simple averaging, leading to suboptimal results in surface reconstruction. In this paper, we focus on the challenging task of *generalizable* 3D human reconstruction and NVS with images from sparse cameras. To this end, we propose the **Di**ffusion-guided Generalizable **Hu**man **NeRF** (DiHuR) framework, which leverages the SMPL model [15] to build geometric-consistent features to infer the 3D structure and incorporates pre-trained diffusion models as geo-

Figure 1. Given 3 views of images with minimal overlap FOV, our method can accurately reconstruct 3D human models.

*Chen Li was at the National University of Singapore when this work was done.

metric guidance to enhance reconstruction quality. Specifically, we fuse the feature by cross-view attention with the learnable tokens attached on the SMPL vertices. The fused features are further refined with several self-attention layers among all the tokens for information exchange. In order to sample the sparse features attached on the fixed number of SMPL vertices, we compute K nearest neighbours and average K features based on the distances. Although effective, the SMPL model is learned from minimal clothes body data and thus fails to model geometric details on the clothes. To solve this, we render the normal map from several target views and feed them to the pre-trained 2D diffusion model to compute SDS loss for fast finetuning. We adapt the diffusion model [20] for super-resolution to provide detailed 3D geometry prior. We first render the normal map from our model, which is then upsampled 4x and used as the input of the diffusion model. The rendered map is also used together with the text as the condition for the denoising process. SDS loss is back-propagated to update our model parameters. A multi-target optimization strategy is also proposed to implicitly regularize the 3D surface. We simultaneously sample rays from different camera views, focusing on the same body part. This ensures intersecting rays between views, which implicitly enforce multi-view consistency during surface prediction. We also apply a second-order Signed Distance Function (SDF) regularization to enhance surface smoothness. Fig 1 shows examples of our accurate 3D reconstruction results on the sparse camera setting. Our main contributions can be summarized as follows:

- We propose DiHuR to solve the challenging task of generalizable 3D human reconstruction and NVS in the sparse views.
- We propose to use learnable latent codes attached on SMPL vertices to guide the sparse view reconstruction and 2D prior from the diffusion model to enhance the details.
- We achieve SOTA-performance for 3D human reconstruction, and novel view synthesis on the commonly used dataset in the sparse view settings.

2. Related works

Human NeRF and Gaussian. NB [18] first combines NeRF with a parametric human body model SMPL [15] to regularize the optimization process. Although it achieves surprising NVS results, it needs long-term optimization and is hard to generalize to unseen identities. Ani-NeRF [17] learns a canonical NeRF model and a backward LBS network which predicts residuals to the deterministic SMPL-based backward LBS (Linear Blending Skinning) to animate the learned human NeRF model. A-NeRF [23] designs a backward mapping with bone-relative representation for feature encoding, which also helps the initial pose correction. Subsequently, generalizable NeRFs are [2,8,25,28] are proposed to make NeRF generalizable to novel scenes in a feed-forward way. They commonly condition the NeRF on the pixel-aligned image features and learn blending weights or directly infer the RGB color condition on such features. However, due to the non-rigid structure and complicated poses of the human body, directly applying such methods fails to work well on real human datasets. [11] utilize view transformers and temporal transformers to aggregate multi-view and multi-frame features to improve the model generalizability. [4] combine the SMPL model and generalizable NeRF in an efficient fashion. However, the 3D reconstruction is far from satisfying due to the lack of 3D constraints, and the NVS is blurry without awareness of the 3D geometry. More recently, people combine 3D gaussian splatting with human template model to reconstruct high quality humans [3, 9, 13]. However, accurate surface extraction from noisy 3DGS remains a challenge.

Human Surface Reconstruction. PIFu [21] is first proposed to reconstruct 3D model from single image with high quality. Subsequently, PIFuHD [22] and ECON [27] is proposed to improve geometric quality with multi-level pixel-aligned features and normal features as additional guidance. More recently, SiTH [7] proposes to combine a back-view hallucination model with an SDF-based mesh reconstruction model. Similarly, SiFU [29] employs a text-to-image diffusion-based prior to generate realistic and consistent textures for invisible views. However, these methods focus on single image reconstruction. Although these methods attempt to hallucinate reasonable interpretations of these invisible areas, the generated regions may not always align with users' expectations. Also, ground truth 3D supervision are used for all of these methods. Compared to these methods, our approach can be trained solely on multi-view images and achieves better performance on sparse view settings.

Diffusion models as guidance. Recently, diffusion models [20] have dominated the image generation area for its exceptional generative quality. Some works [10, 12, 19] therefore utilize diffusion models as guidance to optimize an underlying radiance field for 3D generation and achieve SOTA performance. Specifically, they introduce a Score Distillation Sampling (SDS) loss function, which formulates the same loss function used in the original diffusion model training process, but back-propagate the gradients towards the input rendered images and keeps the pre-trained diffusion models as fixed. With only this SDS loss as supervision, they can generate fantastic 3D meshes with text as an

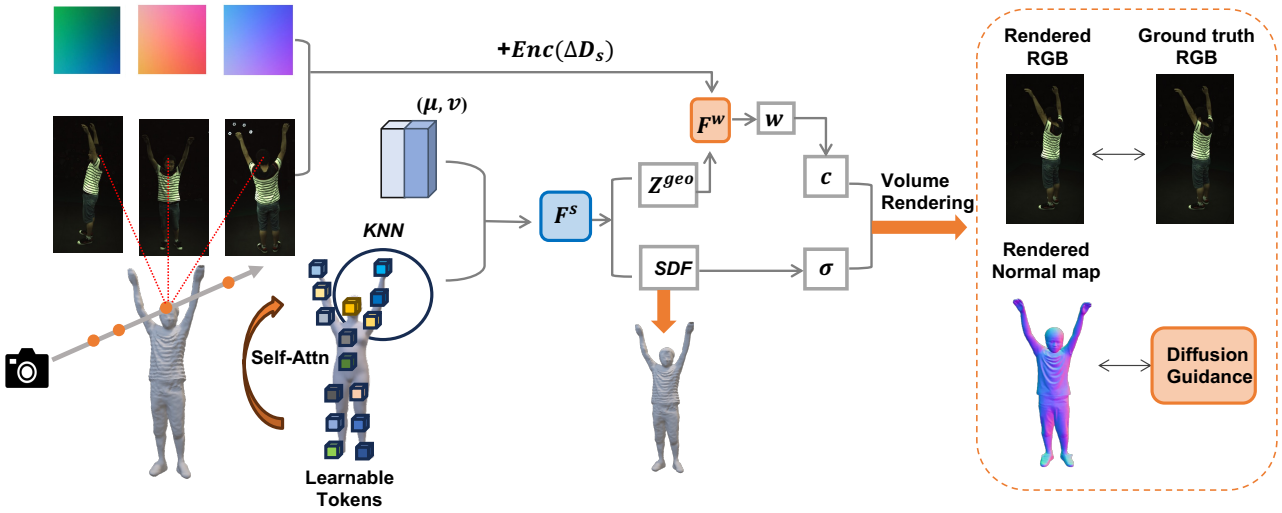

Figure 2. The overview of our proposed DiHuR. Our pipeline mainly consists of three parts: 1) Feature aggregation, 2) SDF prediction, and 3) Appearance prediction. Our learnable tokens serve as query tokens to aggregate sparse view features, which is then interpolated by KNN for each query point. Volume rendering is done with aggregated features and mean, variance of the projected features from sparse input views. During inference, we finetune our model with normal SDS loss to enhance the details.

input condition. However, they often require lengthy optimization due to SDS loss computation and per-scene NeRF optimization from scratch. Inspired by these developments, our approach combines the SDS loss with an initial reconstruction from a generalizable NeRF model. This strategy preserves the prior information from pre-trained diffusion models and leverages the efficiency of feed-forward NeRF models, reducing overall optimization time.

3. Our Method: DiHuR

Problem Analysis. We propose a generalizable reconstruction approach for 3D human reconstruction with sparse cameras. The key challenge in this generalizable setting is how to fuse the features extracted from multi-view images with minimal overlaps, *e.g.* 3 cameras placed 120° apart around the subject. A naive aggregation of the multi-view image features leads to high variance due to inconsistent features from self-occlusions. As a result, the network cannot differentiate between the self-occluded points and points in the free space. Without any supervision on the blending weights, the network can assign colors of occluded image pixels with high blending weights leading to inaccurate color prediction. To circumvent this problem, we propose to set learnable tokens anchored on the SMPL vertices which can learn the common prior across the training dataset to guide the SDF prediction. After the direct inference, we render the normal map from multi-views and finetune the partial parameters with normal SDS loss from the 2D pre-trained diffusion model as geometric guidance to enhance the details.

Overview. As illustrated in Fig. 2, our DiHuR is a volume rendering pipeline with SDF-based volume density representation. We first compute features for each sample point from two sources, *i.e.* directly 3D-to-2D projected features' mean and variance, as well as sparse features aggregated from learnable tokens. These two features are concatenated and then used to predict SDF value along the ray and a geometry code that represents geometry embedding. For color prediction, we use the geometry code and relative ray direction from each of the source camera centers for blending weights prediction. We use ground truth RGB images as supervision. During training, we use multiple target images simultaneously as the supervision to better regularize the underlying 3D geometry. During inference, we finetune our partial model parameters with normal SDS loss.

3.1. Learnable Tokens as Human Prior

Given S source images $\{I_s\}_{s=1}^{S}$ of a human captured by S sparse cameras with minimal overlap FOV, we first use a multi-resolution image feature extractor to get multi-view features $\{F_s\}_{s=1}^{S}$. We use Plucker ray embedding to densely encode the camera poses. The collection of the RGB value and ray embedding for each pixel are concatenated into a 9-channel feature map as the input $I = \{c_i, o_i \times d_i, d_i \mid i = 1, 2, ..., N\}$, where i is the pixel index. In order to aggregate the accurate features from sparse views and make the feature occlusion-aware, we set learnable tokens attached on each of the SMPL vertices as $\{T_q \mid q = 1, 2, ..., 6890\}$. Subsequently, we aggregate the features from S source views for each vertex p_q with a

multi-head cross-view attention module to obtain the SMPL aggregated feature F_q^a as:

$$F_q^a = \sum_{s=1}^{S} w_s^q F_s(p_q) \tag{1}$$

$$w_s^q = SoftMax(\frac{(L_q(T_q)(L_k(F_s(p_q))^\top}{\sqrt{d}}). \tag{2}$$

,where L_q L_k are MLP for query and key embedding, d is the model dimensions. Then, we add self-attention layers to let all the tokens' feature attend each other. Our approach leverages learnable tokens associated with SMPL vertices. These tokens acquire a generalizable prior across diverse identities during training and exploit the consistent mapping of SMPL vertices to similar semantic regions on different human bodies, which serves as a human prior facilitating knowledge transfer to novel identities during inference.

3.2. SDF Prediction with KNN Features

The SMPL feature is sparse in the 3D space since it only contains $6,890$ vertices. When we sample each point on the ray, we first identify K-Nearest Neighbours from all the SMPL vertices and compute the aggregated features based on the distance.

$$Fus(p) = \sum_{k=1}^{K} w_k F_k^a \tag{3}$$

where w_k is computed with an inverse distance function and softmax function. To infer the SDF value for a 3D point p, we project the 3D point to the source image spaces and compute the mean and variance of the image features, *i.e.* $\mu = \frac{1}{S}\sum_{s=1}^{S} F_s(p)$ and v =Var$(\{F_s(p)\}_{s=1}^{S})$. We denote the concatenation of the mean and variance as the global feature $F^{glo}(p) = [\mu, v]$. The global feature is then concatenated with the KNN fused feature at p. The concatenated feature is fed into the SDF prediction network $\mathcal{F}^s$ to predict the SDF value:

$$\{s(p), Z^{geo}\} = \mathcal{F}^s(\gamma(p), [F^{glo}(p), Fus(p)]), \tag{4}$$

where γ denotes the positional encoding, and $s(p)$ denotes the predicted SDF value at p. The network also outputs a geometry code Z^{geo} which can be seen as an implicit local 3D structure encoding. This geometry code is used for the color prediction in the next step.

3.3. Appearance Prediction

Color prediction. We first blend the color among sparse views and then integrate the blended color along the ray to predict the final pixel color. We first compute the difference between two views d and d_s as $\Delta d_s = d - d_s$. We encode this vector to a high dimension as $Enc(\Delta d)$. We then add the view direction feature to the original image features to obtain a new image feature expressed as:

$$F_s^{\oplus} = F_s + Enc(\Delta d). \tag{5}$$

Then, we concatenate image feature $F_s^{\oplus}$, the geometry code Z^{geo} and feed it into our blending weights prediction network $\mathcal{F}^w$ for view blending weights w_s prediction:

$$w_s = \mathcal{F}^w(F_s^{\oplus}, Z^{geo}). \tag{6}$$

This information contains rich geometry clues to help the network infer the correct blending weight for the source images with few color correspondences due to view sparsity. The color for each 3D point is the weighted sum of the projected color $c_s = I_s(\pi_s(p)$ from S source images, *i.e.*:

$$c = \sum_{s=1}^{S} w_s c_s. \tag{7}$$

Volume rendering. For volume rendering along the rays, we follow the NeuS [24] formulation to convert the SDF value $F^s(p)$ to the density σ , *i.e.*:

$$\sigma = max(-\frac{\frac{d\Phi}{dp}(\mathcal{F}^s(p))}{\Phi(\mathcal{F}^s(p))}, 0), \tag{8}$$

where p is the point sampled on the ray and Φ is the Sigmoid function. Finally, for a ray sampled with M points, the color is accumulated as:

$$\begin{aligned} C(r) &= \sum_{i=1}^{M} T_i(1 - exp(-\sigma_i))_i, \\ \text{where } T_i &= exp(-\sum_{j=1}^{i-1} \sigma_j). \end{aligned} \tag{9}$$

3.4. Score Distillation Sampling for Geometry Enhancement

During the inference, as shown in Fig. 3, we compute the $\mathcal{L}_{sds}$ [19] with an pre-trained diffusion model. Different from the original SDS loss in [19], we choose to use a normal map as the input to the super-resolution diffusion model in order to enhance the geometric details. More specifically, we use our pretrained model to infer the normal map as the low-resolution image condition. We then fine-tune our model with the guidance from the diffusion model. Note that we only finetune $\mathcal{F}_s$ for fast convergence. We rendered the normal map and up-sample $4\times$ for this normal map and subsequently get the latent code from the VAE encoder in the diffusion model. The normal $N(r)$ for each ray is computed by volume rendering with the gradient of the points'

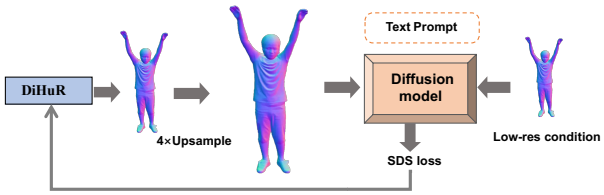

Figure 3. We show how to compute SDS loss on the pre-trained super-resolution diffusion model: the low-resolution conditioned image on the right is generated from our fixed model.

SDF sampled on each ray as:

$$N(r) = \sum_{i=1}^{M} T_i(1 - exp(-\sigma_i))\frac{\nabla \mathcal{S}(p_i)}{||\nabla \mathcal{S}(p_i)||}, \quad \text{where } T_i = exp(-\sum_{j=1}^{i-1}\sigma_j), \tag{10}$$

We finally compute the SDS loss:

$$\mathcal{L}_{sds} = w(t)(\hat{\epsilon}_\phi(z_t; y, N^{low}, t) - \epsilon)\frac{\partial z_t}{\partial N}\frac{\partial N}{\partial \theta}, \tag{11}$$

with the input text and low-resolution normal map rendered from our fixed model as conditions. z_t is the noisy latent by adding noise to the latent from the up-sampled rendered normal map N, and θ are our model parameters. $\hat{\epsilon}_\phi$ is the noise prediction model and ϵ is the random noise added, t is sampled from 0.52 to 0.98 empirically, N^{low} is the low-resolution normal map which is detached during the optimization. y is the text condition, we set it as 'Best quality, human, normal map'. For classifier-free guidance(CFG), we set the guidance weights as 7.5. The intuition is to provide detailed normal supervision for our pre-trained network based on the 2D image super-resolution diffusion model. During SDS refinement, we select 8 views with azimuth evenly increasing from 0 to 360 degree to render the normal map. We also add image reconstruction loss for the input views as well as other regularization losses. The whole process takes 150 iterations around 2 minutes.

3.5. Optimization

Multi-target optimization. We believe multi-view images can serve as the surrogate for 3D supervision when direct supervision of the SDF values is not available. Consequently, we use multiple images from different views as the target in each iteration instead of using one. The random sampling strategy used in the original NeRF can lead to non-intersecting rays between the multiple target views. To solve this problem, we sample patches from the same body part in each target image to concentrate the optimization zone. Specifically, we simultaneously sample patches from the same segmentation part, *e.g.* upper body, in the multiple target images at each iteration. This sampling strategy results in more overlapping points from different views, hence providing more explicit multi-view supervision. The loss of our multi-target optimization is given by:

$$\mathcal{L}_{rgb} = \sum_{l=1}^{L}\sum_{r\in R} ||\hat{C}(r)_l - C(r)_l||. \tag{12}$$

L is the number of target images for ray sampling and R is the set of sampled rays for each target image.

Total loss. We enforce the normal vectors of nearby points to be similar with a smooth regularization, *i.e.*:

$$\mathcal{L}_{sm} = \sum_{p\in S} ||\nabla \mathcal{S}(p) - \nabla \mathcal{S}(p+\epsilon)||_2^2, \tag{13}$$

where S represents the set of sampled points, and ϵ represents a perturbation sampled from a Gaussian distribution. We also include Eikonal regularization, *i.e.*:

$$\mathcal{L}_{eik} = \sum_{p\in S}(||\nabla \mathcal{F}^s(p)||_2 - 1)^2. \tag{14}$$

Our full objective function contains the color loss $\mathcal{L}_{rgb}$, SDS loss, the Eikonal regularization [6] $\mathcal{L}_{eik}$, second-order SDF regularization $\mathcal{L}_{sm}$ with corresponding weights as hyper-parameters, *i.e.*:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{sds}\mathcal{L}_{sds} + \lambda_{eik}\mathcal{L}_{eik} + \lambda_{sm}\mathcal{L}_{sm}. \tag{15}$$

Note that only during the per-scene finetuning stage, we add SDS loss.

4. Experiments

4.1. Experiment Setup

Dataset. We conduct experiments on the commonly used ZJU-MoCap dataset [18], THuman dataset [30] and HuMMan dataset [1]. THUman contains 202 human body 3D scans. Following [4], 80% of the scans are taken as the training set, and the remaining are the test set. For ZJU-MoCap dataset, it consists of 9 sequences captured with 23 calibrated cameras. Following [4], we train our model using 6 sequences and test our model on the remaining 3 sequences for 3D reconstruction. We further train our method on THuman dataset and conduct cross-dataset evaluation on HuMMan [1] dataset. Specifically, we use the last 22 sequences for evaluation.

4.2. 3D Reconstruction

Baselines. Baselines: we compare our method with Generalizable NeRF methods: MPS-NeRF [5], SparseNeuS [14] and GP-NeRF [4] as well as single view human reconstruction methods PIFuHD [22], SiTH [7] and SIFU [29].

Figure 4. From left to right are reconstruction results from NeuS [24], PIFuHD [22], SparseNeuS [14], GP-NeRF [4], SiTH [7], SIFU [29], Ours, and ground truth. For single view reconstruction methods [7, 22, 29], we choose the front view as input.

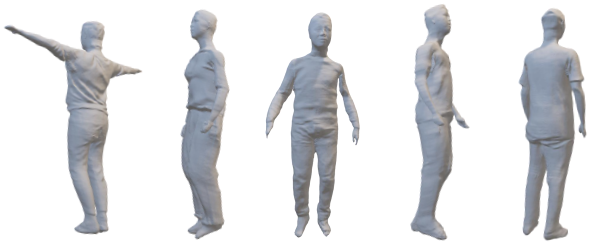

Figure 5. Zero-shot reconstruction results on HuMMan dataset.

On ZJU-MoCap dataset, we also compared with NeuS [24] and NB [18]. Following [4], we use 3 source views with minimal overlap FOV as the input. For single view reconstruction methods, we choose the front view as input. We use commonly used Chamfer Distance (CD) and Normal Consistency (NC) as the metrics for evaluation.

Evaluation. We show our 3D reconstruction results with other existing methods in Tab. 1 for THuman dataset and Tab. 3 for ZJU-MoCap. All the NeRF-based models are trained and tested on the same splits without 3D supervision. Our methods outperform existing methods [4, 5, 14, 18, 24, 29] with a large margin in terms of both CD and NC. We show the qualitative comparison with existing methods in Fig. 4. We can see in the figure that generalizable NeRF methods [4, 5, 14] fail to reconstruct the details although the pose and shape are correct. Directly optimizing NeuS in this setting produces large noisy areas with disconnected points as shown in Fig. 4. PIFuHD also shows poor generalization ability on this dataset with missing arms and legs as shown in Fig. 4 with extremely large error, and therefore we do not include it in the quantitative results. SiTH and SIFU tend to reconstruct smooth meshes with wrong poses due to single view ambiguity, e.g. the second row in Fig. 4 compared with GT. In contrast, our model can infer a more realistic human mesh with fine de-

tails such as wrinkles on the clothes as well as correct body shape and pose. We also show cross-dataset results in Fig.5 to demonstrate the generalization ability, where our model is trained on ZJU-MoCap dataset. Note for ZJU-MoCap dataset, we use all the images and NeUS [24] to obtain the GT 3D meshes. For SMPL we use

Methods	[14]	[5]	[4]	[7]	[29]	Ours
CD(↓)	2.312	3.312	3.876	1.621	1.521	**1.117**
NC(↑)	0.634	0.616	0.567	0.723	0.741	**0.779**

Table 1. 3D reconstruction comparison on THuman [30] dataset.

Methods	[14]	[5]	[4]	[7]	[29]	Ours
CD(↓)	3.211	3.412	3.765	1.723	1.709	**1.23**
NC(↑)	0.633	0.621	0.612	0.711	0.721	**0.753**

Table 2. Cross-dataset evaluation on HuMMan [1] dataset.

Methods	[18]	[24]	[14]	[4]	[7]	[29]	Ours
CD(↓)	1.876	4.544	5.875	2.448	1.123	0.942	**0.790**
ND(↑)	0.624	0.478	0.403	0.578	0.716	0.707	**0.767**

Table 3. 3D reconstruction comparison on ZJU-MoCap dataset.

Method	P.S.	U.B.	U.P.	PSNR(↑)	SSIM(↑)
Seen people on seen frames					
NB [18]	✓	✗	✗	28.51	0.947
MPS-NeRF [5]	✗	✓	✓	28.54	0.933
NHP [11]	✗	✗	✗	28.73	0.936
GP-NeRF [4]	✗	✗	✗	28.81	0.944
Ours	✗	✗	✗	**28.94**	**0.951**
Seen people on unseen frames					
NB [18]	✓	✗	✓	23.79	0.887
MPS-NeRF [5]	✗	✓	✓	27.02	0.931
NHP [11]	✗	✗	✓	26.94	0.929
GP-NeRF [4]	✗	✗	✓	27.92	0.934
Ours	✗	✗	✓	**28.35**	**0.941**
Unseen people on unseen frames					
NB [18]	✓	✓	✗	22.88	0.883
MPS-NeRF [5]	✗	✓	✓	25.17	0.911
NHP [11]	✗	✓	✓	24.75	0.906
GP-NeRF [4]	✗	✓	✓	25.96	0.921
Ours	✗	✓	✓	**26.26**	**0.926**

Table 4. We thoroughly compared our method with existing NVS works under three settings on ZJU-MoCap dataset. P.S. refers to per-scene optimization, U.B. refers to unseen bodies, and U.P. refers to unseen poses.

4.3. Image Synthesis

Baselines. For ZJU-MoCap, we show results of novel view synthesis in 3 source views with minimal overlap FOV setting and compare with existing NeRF-based methods [4,11,18]. For THuman dataset, we follow [5] and compare with existing human NeRF methods [4, 5, 17, 18]. We use commonly used SSIM and PSNR as evaluation metrics.

Methods	NB [18]	NHP [11]	MPS [5]	GP-NeRF [4]	Ours
SSIM(↑)	0.907	0.895	0.914	0.923	**0.931**
PSNR(↑)	24.86	24.10	24.63	24.88	**26.31**

Table 5. NVS comparison on THuman dataset.

Evaluation. For ZJU-MoCap dataset, following GP-NeRF [4], we show the generalization from three aspects: novel view synthesis on the seen bodies with seen poses, seen bodies with unseen poses, and unseen bodies. We show the results in Tab. 4 compared with the generalizable human novel view synthesis methods. As shown in the table, we achieve the best performance than all the other methods. For THuman dataset, we show the quantitative comparison with SOTA methods in Tab. 5. We only test the unseen bodies on this dataset and achieve the best performance. We credit this to our more accurate 3D geometry modeling with SDS guidance and SMPL localized features. We show the qualitative results in Fig. 6 only with [4], which demonstrated the highest performance among all other methods in our evaluation in Tab.4. Compared with GP-NeRF [4], our method can synthesize images with better color and finer details. As highlighted in the red box of Fig. 6, GP-NeRF generates wrong colors which are not consistent with the input images for some areas while the colors generated from our method show more consistency. While our method can synthesize images with more details. This is because with learned tokens as guidance, it can predict the accurate blending weights learned from the training data, which greatly reduces the blurry effect. Also, SDS optimization also helps color prediction since our blending weights prediction is relied on the accurate geometry prediction.

4.4. Ablation studies.

We show the ablation studies for 3D reconstruction and novel view synthesis on THuman dataset in terms of learnable tokens, diffusion guidance, multi-target optimization in Tab.7. As shown in the table, we can see that diffusion guidance plays an important role in both 3D reconstruction and novel view synthesis because our whole network is only supervised with images, and the whole network can converge better with the geometric SDS loss as guidance. Our learnable tokens also significantly improve both 3D and 2D

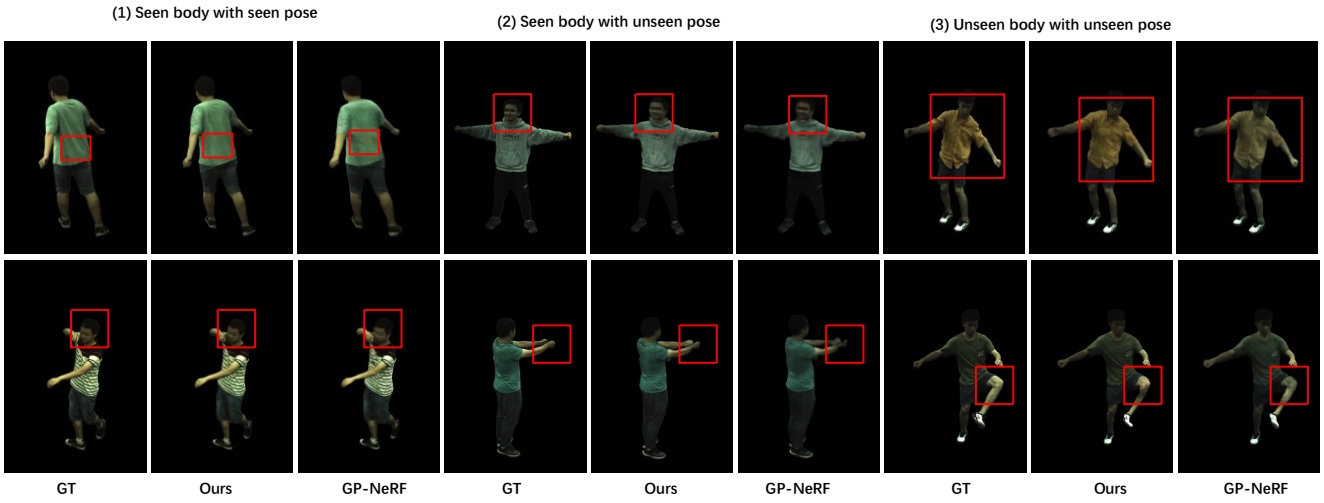

Figure 6. Visual comparison with GP-NeRF [4]. From left to right are the results from: 1) Seen body with seen pose; 2) Seen body with unseen pose; 3) Unseen body with unseen pose. We show two examples for each setting. The details are highlighted in the red boxes.

results. Learnable tokens provides vertex-to-semantic mapping enables our model to generalize well to unseen subjects. Multi-target training makes the 3D geometry even better both quantitatively and qualitatively. We also show the reconstruction results with different number of input views in Tab.6. Increasing the number of input views, both NVS and 3D reconstruction quality are improved.

Number of views	NVS		3D geometry	
	PSNR(↑)	SSIM(↑)	CD(↓)	NC(↑)
2	24.98	0.913	1.345	0.702
3	26.31	0.931	1.117	0.779
4	26.88	0.938	1.011	0.787
5	**27.07**	**0.942**	**0.987**	**0.792**

Table 6. Ablation study for number of input views.

Components	NVS		3D geometry	
	PSNR(↑)	SSIM(↑)	CD(↓)	NC(↑)
w/o learnable code	23.98	0.907	1.543	0.667
w/o diffusion guidance	25.58	0.921	1.344	0.713
w/o multi-target training	26.08	0.926	1.243	0.757
full model	**26.31**	**0.931**	**1.117**	**0.779**

Table 7. Ablation study for each component.

We also show the visual comparison in Fig. 7 to validate the effectiveness of each component. We can see from the figure that without the SDS guidance, the 3D reconstruction suffers from losing details in both face and clothes, which is also the case for the results without learnable tokens as feature aggregation guidance. Multi-target optimization constraint compresses noisy part to make our reconstruction

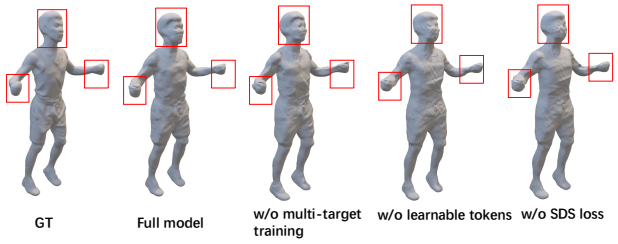

Figure 7. Ablation studies on each component in our proposed method. See the detailed comparison highlighted in the red boxes.

looks more natural. Details in the arms and facial part are highlighted in the red box. Our full model achieves precise 3D reconstruction, capturing fine details with high accuracy.

5. Conclusion

In this paper, we introduce DiHuR for generalizable 3D human reconstruction and novel view synthesis from sparse cameras. To mitigate the minimal overlapping view problem, we introduce learnable tokens attached on human parametric model to guide the sparse view feature aggregation and SDF prediction process. Without 3D supervision, we utilize a 2D pre-trained diffusion model as the normal guidance to improve geometry details. We further propose a multi-target training strategy to constrain the underlying 3D surface. Quantitative and qualitative results on the commonly used benchmark show the superiority of our DiHuR compared with existing approaches.

References

[1] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 5, 7

[2] 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 *CVPR*, 2021. 2

[3] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024. 2

[4] 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 *ECCV*, 2022. 1, 2, 5, 6, 7, 8

[5] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *PAMI*, 2022. 1, 5, 6, 7

[6] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 5

[7] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5, 6, 7

[8] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. *arXiv preprint*, 2023. 1, 2

[9] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos, 2023. 2

[10] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2

[11] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *NeurIPS*, 2021. 2, 7

[12] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[13] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023. 2

[14] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, 2022. 5, 6, 7

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. In *ICCV*, 2015. 1, 2

[16] 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*, 2020. 1

[17] 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*, 2021. 2, 7

[18] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1, 2, 5, 6, 7

[19] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2, 4

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[21] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

[22] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2, 5, 6

[23] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021. 1, 2

[24] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 4, 6, 7

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

[26] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, 2022. 1

[27] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 512–523, June 2023. 2

[28] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[29] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9936–9947, June 2024. 2, 5, 6, 7

[30] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, 2019. 5, 7