Supplementary Material for DiHuR: Diffusion-Guided Generalizable Human Reconstruction

We provide supplementary material in terms of more training and inference details, cross dataset evaluation, and more quantitative and qualitative comparison on THuman dataset, and more visual results rendered at the view different from input views of the reconstructed meshes.

Ablation study for K.

We ablates the number of Nearest Neighbours: K on THuman dataset in Tab.1. Smaller K leads to insufficient information, but larger K also includes confusing information from other long range body parts. K is chosen to be 5 in our main experiments according to the experiments.

Number of views	NVS		3D geometry	
	PSNR(↑)	SSIM(↑)	CD(↓)	NC(↑)
5	25.98	0.913	1.345	0.721
10	**26.31**	**0.931**	**1.117**	**0.779**
15	26.27	0.925	1.211	0.732
20	26.11	0.915	1.234	0.712

Table 1. Ablation study for number of nearest neighbors.

More comparison, training and inference details. The hyper-parameters in our loss function are set as follows: $\lambda_{rgb} = 1$ for the RGB reconstruction loss, $\lambda_{eik} = 0.01$ for the Eikonal regularization term, $\lambda_{sds} = 0.1$ for the signed distance surface loss, and $\lambda_{sm} = 0.01$ for the smoothness constraint. We add an additional background loss term to suppress noise in background areas:

$$\mathcal{L}bg = \frac{1}{B} \sum r \in B||W(r)||, \tag{1}$$

where $W(r)$ denotes the accumulated weights along the ray for pixel r in the background region B. This loss term is weighted by $\lambda_{bg} = 0.01$ in the total loss function. For efficient ray sampling, we adopt a hierarchical strategy with reduced sampling points per ray. Our approach first uniformly sampling 8 points along each ray, followed by four iterations of importance sampling based on the SDF, with 12 points sampled in each iteration. This strategy ensures adequate sampling density while maintaining computational efficiency. The network architecture is configured as follows: The multi-head attention module uses 4 layers with 4 attention heads each. For processing input images, we extract features with a dimension of 32. The SDF prediction network comprises 4 linear layers augmented with position embedding to enhance spatial understanding. To ensure fair comparison across methods, we utilize ground truth SMPL parameters for all evaluated approaches.

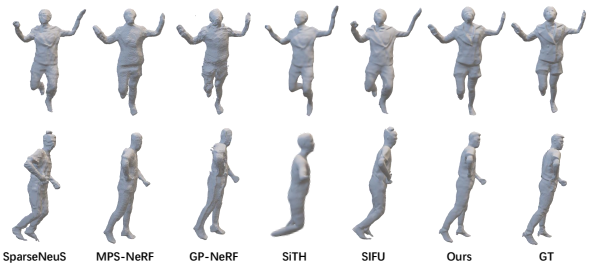

Figure 1. Visual comparison on THuman dataset. From left to right are results from SparseNeuS [5], MPS-NeRF [2], GP-NeRF [1], SiTH [3], SIFU [6], Ours and ground truth.

More visual results We provide visual comparison of our method with other existing methods on the THuman dataset in Fig. 1. As shown in the figure, our method demonstrates clear advantages over previous approaches in better preserving geometric details including clothing wrinkles and other fine structures.

We provide more visualizations on ZJU-MoCap and THuman, with additional results shown in Fig.2 and Fig.3 respectively. The comprehensive evaluation demonstrates our method's consistent performance across different subjects and poses. Moreover, to test the generalization capability of our model, we perform zero-shot cross-dataset evaluation on the CustomHumans [4] dataset using higher quality input images, as presented in Fig. 4. For each subject, we render novel views that differ from the input viewpoints. The results clearly demonstrate our model's strong generalization ability to unseen subjects. This robust cross-dataset performance validates the effectiveness of our approach in real-world applications.

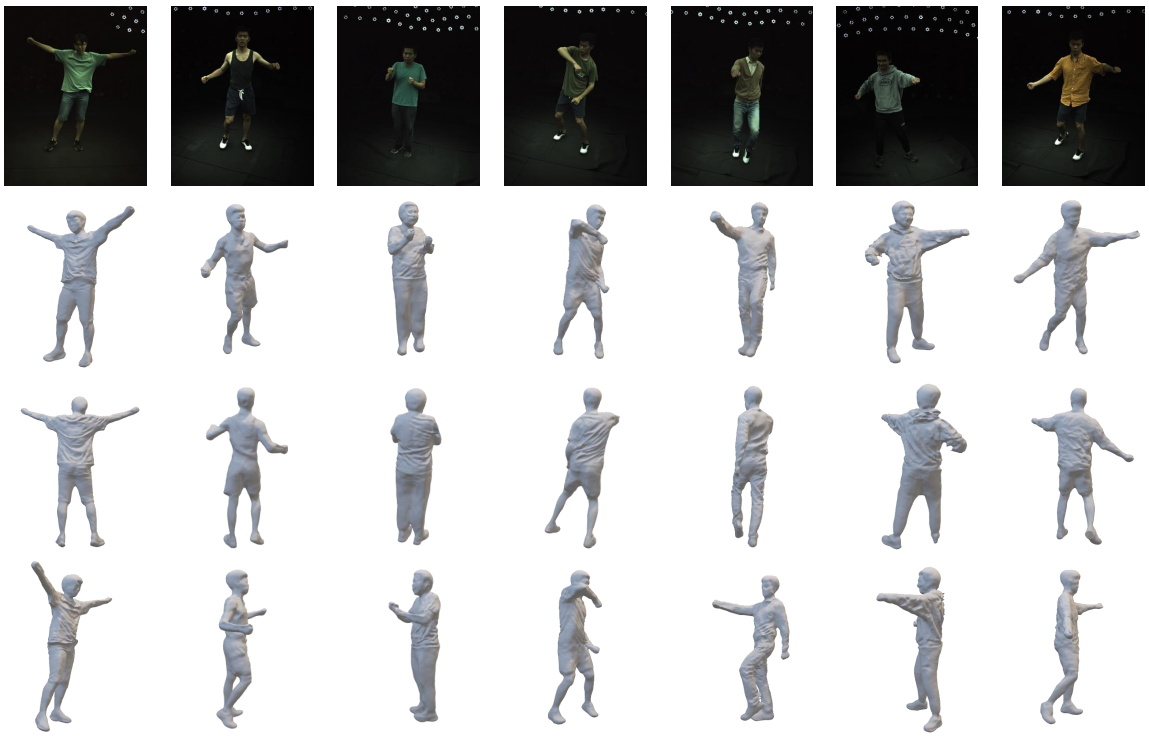

Figure 2. 3D reconstruction for ZJU-MoCap dataset.

Figure 3. 3D reconstruction for THuman dataset.

Figure 4. Zero Shot 3D reconstruction for CustomHuman dataset.

References

[1] 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] 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

[3] 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. 1

[4] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[5] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, 2022. 1

[6] 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. 1