


# Supplemental Document for MeshAvatar: Learning High-quality Triangular Human Avatars from Multi-view Videos

Yushuo Chen<sup>1</sup>, Zerong Zheng<sup>2</sup>, Zhe Li<sup>1</sup>, Chao Xu<sup>2</sup>, and Yebin Liu<sup>1</sup>

<sup>1</sup> Tsinghua University, Beijing, China

<sup>2</sup> NNKosmos Technology, Hangzhou, China

In this supplemental document, we present more details about our implementation & experiments (Sec. A), show more results and additional experiments (Sec. B), and discuss several potential social impacts (Sec. C). More visual results are demonstrated in the supplemental video.

## A Implementation Details

Our learning objective

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{reg}} \quad (1)$$

consists of the photometric loss between rendered image  $\hat{\mathbf{I}}$ /normal map  $\hat{\mathbf{N}}$  and the input image  $\mathbf{I}$ /estimated normal  $\mathbf{N}$ , plus the regularization terms for our implicit SDFs, pose-dependent vertex offsets, materials and lighting.

$$\mathcal{L}_{\text{img}} = \left\| \hat{\mathbf{I}} - \gamma_{\text{tone}}^{-1}(\mathbf{I}) \right\|_1 + \left\| \hat{\mathbf{N}} - \mathbf{N} \right\|_1 + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}} \left( \gamma_{\text{tone}}(\hat{\mathbf{I}}), \mathbf{I} \right), \quad (2)$$

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{SDF}} \mathcal{L}_{\text{SDF}} + \lambda_{\text{offset}} \mathcal{L}_{\text{offset}} + \lambda_{\text{mat}} \mathcal{L}_{\text{mat}} + \lambda_{\text{light}} \mathcal{L}_{\text{light}}, \quad (3)$$

where  $\mathcal{L}_{\text{LPIPS}}$  is the perceptual loss [12],  $\gamma_{\text{tone}}$  is the tone mapping function to map the rendered image from linear color space to sRGB color space, and  $\lambda_{\text{LPIPS}}, \lambda_{\text{SDF}}, \lambda_{\text{mat}}, \lambda_{\text{offset}}, \lambda_{\text{light}}$  are balancing coefficients,

$$\mathcal{L}_{\text{SDF}} = \sum_{x \in \mathbb{R}^3} \left\| \nabla_x \mathcal{S}(x) - 1 \right\|^2, \quad \mathcal{L}_{\text{offset}} = \frac{1}{N_c} \sum_{n=1}^{N_c} \left\| \Delta \mathbf{v}_n \right\|^2 \quad (4)$$

encourages the base mesh to be smooth [2] and fit the clothed human as much as possible, and

$$\mathcal{L}_{\text{mat}} = \sum_{\mathbf{v} \in \mathcal{M}, k \in \{k_d, k_s\}} |k(\mathbf{v}) - k(\mathbf{v} + \epsilon)|, \quad (5)$$

$$\mathcal{L}_{\text{light}} = \frac{1}{3} \sum_{c \in \{r, g, b\}} \left| L_{\text{env}, c} - \frac{1}{3} \sum_{c \in \{r, g, b\}} L_{\text{env}, c} \right| \quad (6)$$

are used to regularize spatially-smooth material and low-frequency lighting [7].

**Table A:** Comparisons on training/inference time with other SOTA methods on neural avatars.

	Ours	PoseVocab [4]	AvatarReX [14]	Animatable Gaussians [5]	Xu <i>et al.</i> [10]	Lin <i>et al.</i> [6]	Wang <i>et al.</i> [9]
Relightable?	✓				✓	✓	✓
Training Time (~100 frames)	~3h					2.5 days	4h (monocular)
Training Time (~1000 frames)	~16h	1.5~2 days	2 days	2 days (RTX 4090)	30h		
Inference Time (per image)	180ms	3s	30s	100ms (1024×1024)	5s	40s	20s

Besides dataset preprocessing like SMPL-X registration and stereo-based normal estimation, our avatar modeling pipeline is completely end-to-end, with the supervision signal from Equation 1. To ensure that the geometric details can be generated, the resolution of our tetrahedral grid is set 256 (Every edge in the grid has length 1/128 m). The tetrahedral grid is mainly defined around the SMPL-X body shape to avoid SDF queries in unnecessary regions. The extracted human mesh has  $\sim 35k$  vertices and  $\sim 70k$  triangles. The SDF field is implemented and initialized the same as in VolSDF [11]. During training, the balancing loss coefficients are set  $\lambda_{\text{LPIPS}} = 0.1$ ,  $\lambda_{\text{SDF}} = 0.01$ ,  $\lambda_{\text{mat}} = 0.3$  for diffuse albedo and 0.05 for surface roughness.  $\lambda_{\text{offset}}$  decreases linearly from 1000 to 10 to learn a meaningful base mesh at the early steps.

Our model is trained on single NVIDIA RTX 3090 GPU using Adam [3] optimizer. It takes 100k steps and around 16 hours to converge. At the inference time, we sample 64 reflective rays without denoiser for more realistic and accurate rendering. It takes only 180ms (35ms to generate posed mesh + 145ms to render) to render an image of  $512 \times 512$  resolution, in contrast to 40s in [6],  $\sim 20s$  in [9], 5s in [10] and 50s in [1], proving the effectiveness of our mesh-based representation. More comparisons with other SOTA neural avatars are shown in Table A. Existing works using 3DGS [5] could not produce accurate dynamic geometries nor support relighting under novel environments. Note that the former 35ms is inevitable to produce pose-dependent dynamic details using neural networks (even if using 3DGS), and the latter 145ms was measured using the same differentiable Monte-Carlo renderer as in training. Given that the textured mesh has been obtained in the former step, the rendering time can be significantly reduced by advanced rendering techniques (e.g. NVIDIA RTX), making real-time rendering feasible in the future.

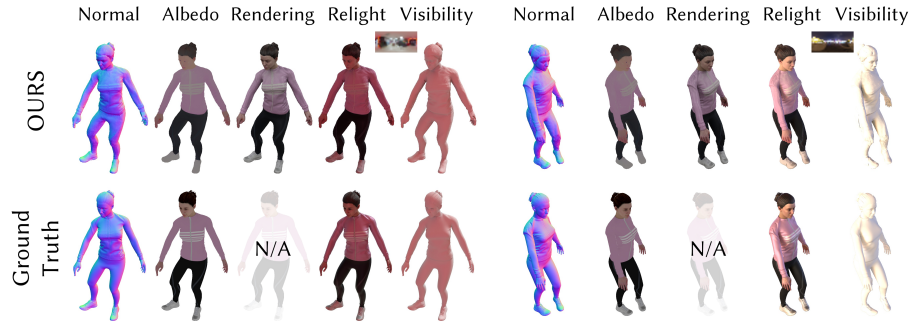
## B More Validations on Inverse Rendering and Relighting

### B.1 Evaluations on Synthetic Data

For the task of intrinsic decomposition, we further evaluate our method on a synthetic dataset SyntheticHuman++ [8] and compare with state-of-the-art methods [1, 10, 13]. This dataset consists of 100 frames  $\times$  20 camera views, from

**Table B:** Quantitative Comparisons on *SyntheticHuman++* dataset. Following [10], Normal degree and PSNR are computed only in the pixels with foreground mask activated, while SSIM and LPIPS are computed in the bounding box of the human region. The metric computations are slightly different to those in the main text. The best results and the second best are highlighted in **bold** and underlined fonts, respectively.

	Normal	Albedo			Relighting			Visibility		
	Degree↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours	<u>13.44</u>	<b>31.94</b>	<b>0.953</b>	<b>0.073</b>	<b>27.19</b>	<b>0.941</b>	<b>0.066</b>	<b>22.30</b>	<b>0.910</b>	<b>0.086</b>
Ours (w/o normal)	19.37	<u>31.08</u>	<u>0.942</u>	<u>0.072</u>	<u>26.92</u>	<u>0.939</u>	<u>0.068</u>	<u>20.36</u>	<u>0.891</u>	<u>0.098</u>
Xu <i>et al.</i> [10]	<b>12.44</b>	29.01	0.933	0.119	22.69	0.861	0.206	20.20	0.848	0.155
Relighting4D [1]	29.38	24.70	0.885	0.183	22.13	0.835	0.237	15.22	0.763	0.252
NeRFactor [13]	34.29	22.23	0.817	0.226	21.04	0.758	0.313	11.37	0.581	0.387



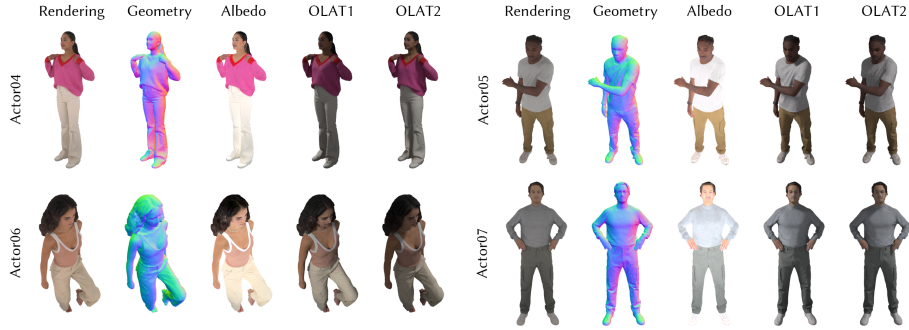
**Fig. A:** Visualizations of our method on character *jody* in *SyntheticHuman++* dataset.

which we use 10 for training and the others for novel view testing. We use this original training/testing split in the dataset for evaluation, and follow the same protocol in [10]. The results of [1, 10, 13] are borrowed from [10]. Due to the scale ambiguity of inverse rendering problem, the metrics are computed after scale alignment [6, 9, 10] on each color channel. As demonstrated in Table B, our method significantly outperforms Relighting4D [1] across all inverse rendering metrics. Furthermore, benefiting from explicit mesh representation and more accurate rendering, our method also achieved notable improvement from Xu *et al.* [10] on the quality of novel light synthesis, despite limited enhancements in geometry reconstruction. Example qualitative results of our method are shown in Figure A.

We also evaluated the effectiveness of pseudo normal supervision on this synthetic dataset. Tab. B presents the quantitative result without using estimated normals, denoted as *Ours (w/o normal)*. In comparison with *Ours*, it demonstrates that introducing geometric priors enhances the accuracy of geometry reconstruction, which subsequently improves image synthesis under novel lighting. We still achieved better performance than other SOTA methods in most metrics, except geometric error, underscoring the efficacy of our proposed method.

**Table C:** Additional quantitative comparisons on reconstructions and novel pose synthesis. The better one is highlighted in **bold** fonts.

	Training Frame Reconstruction				Novel Pose Synthesis			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
AvatarReX	28.4015	0.9607	0.0579	30.6214	26.5853	<b>0.9497</b>	<b>0.0684</b>	36.1911
Ours (w/o normal)	<b>30.2084</b>	<b>0.9616</b>	<b>0.0564</b>	<b>24.7444</b>	<b>27.1589</b>	0.9495	0.0687	<b>31.2888</b>

**Fig. B:** Visualizations of our learned avatars synthesized under OLAT environments.

## B.2 Additional Experiments

**Relighting under OLAT.** To further demonstrate the solved intrinsic properties and the relighting capability of our method, we relight our learned avatar using OLAT (One Light At a Time) environment maps and visualize them in Figure B. As shown in this figure, our method is able to realistically synthesize the shading effects of cloth wrinkles under different lighting directions. This experiment further proves that our method is able to recovery accurate geometry and albedo/material for dynamic humans.

**Comparisons without Normal Priors.** Considering our method employed additional priors from pseudo normal supervision, which may bias the comparison in the main text, we further report the quantitative result without using estimated normals in Tab. C. The evaluation is performed on AvatarReX dataset and the metrics are the same as we used in Tab. 1 in the main text.

**Correspondences.** Another advantage of our method is that it naturally realizes surface tracking and establishes point-to-point correspondence among the whole performance sequence, which is typically difficult, if not impossible, in previous implicit representations. Figure C shows the color-coded correspondences across different poses. The rendered colors on the right image of each sub-figure are defined as the corresponding normalized canonical coordinates of the ray-traced points. It demonstrates that our method learns reasonable mesh correspondences from images without explicit surface tracking.





**Fig. C:** Visualizations of the correspondences of our learned avatar.

## C Potential Social Impacts

Our method facilitates the automatic digital image creation of any specific human identity. However, this capability poses the risk to generate fake motion sequences that the individual has never performed. This issue should be carefully addressed before deployment.

## References

1. Chen, Z., Liu, Z.: Relighting4d: Neural relightable human from videos. In: European Conference on Computer Vision. pp. 606–623. Springer (2022)
2. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: ICML. pp. 3789–3799. PMLR (2020)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
4. Li, Z., Zheng, Z., Liu, Y., Zhou, B., Liu, Y.: Posevocab: Learning joint-structured pose embeddings for human avatar modeling. In: ACM SIGGRAPH Conference Proceedings (2023)
5. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. arXiv preprint arXiv:2311.16096 (2023)
6. Lin, W., Zheng, C., Yong, J.H., Xu, F.: Relightable and animatable neural avatars from videos. AAAI (2024)
7. Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting Triangular 3D Models, Materials, and Lighting From Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8280–8290 (June 2022)
8. Peng, S., Dong, J., Wang, Q., Zhang, S., Shuai, Q., Zhou, X., Bao, H.: Animatable neural radiance fields for modeling dynamic human bodies. In: ICCV. pp. 14314–14323 (2021)
9. Wang, S., Antić, B., Geiger, A., Tang, S.: Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
10. Xu, Z., Peng, S., Geng, C., Mou, L., Yan, Z., Sun, J., Bao, H., Zhou, X.: Relightable and animatable neural avatar from sparse-view video. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2024)
11. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. NeurIPS **34**, 4805–4815 (2021)

12. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
13. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)* **40**(6), 1–18 (2021)
14. Zheng, Z., Zhao, X., Zhang, H., Liu, B., Liu, Y.: Avatarrex: Real-time expressive full-body avatars. *TOG* **42**(4) (2023). <https://doi.org/10.1145/3592101>