

Illumination-Consistent Human-Scene Reconstruction from Monocular Video

Supplementary Material

In this supplementary material, we first provide additional implementation details to complement the method described in the main manuscript in Sec. A. Sec. B introduces the architecture of the light volume and the decoder. Sec. C provides more comparison results, including scene reconstruction on the NeuMan dataset, human reconstruction on the ZJU-MoCap dataset, and runtime analysis of shadow estimation. Sec. D provides additional ablation study results on our two-stage human reconstruction strategy and the depth constraint. Finally, to demonstrate the effectiveness of our method in reconstructing human, scene lighting and shadows, Sec. E presents additional relighting results, human scene transfer outcomes, and visualizations of the intermediate outputs in our pipeline.

A. Implementation Details

Our implementation is based on PyTorch, utilizing the Adam optimizer for training. All experiments are performed on a single NVIDIA 4090 GPU, and the training process requires 22GB of GPU memory and takes approximately 2.5 hours.

A.1. More Details of Human Reconstruction

In the first stage, we up-sample the SMPL to a new mesh with 27.5k vertices and 55.1k faces. We use the centroids of the triangular faces as the positions for the Gaussians and determine each Gaussian’s rotation by aligning the face normal with its shortest axis. The opacity attribute is defined as learnable parameters. We optimize the hash encoder to obtain the positional offsets of the mesh vertices and indirectly refine the face Gaussians, which are not densified/pruned. After training without PBR for 4,000 iterations, the human mesh is fixed.

In the second stage, we no longer optimize the face Gaussians in the first stage; they are only used to initialize the human Gaussians at the beginning of the second stage. We clone, split, and prune Gaussians every 200 iterations using the KL divergence [4]. The optimization takes 15k iterations, and the maximum number of human Gaussians is about 200k. To ensure that the light probes cover the human body, PBR is skipped for the first 1000 iterations to track the human’s position and align the light volume to the occupied region. During the first 1000 iterations, we focus on learning the human base color. This strategy helps to expedite the disentanglement of lighting and human appearance and prevents human appearance from influencing the learning of scene illumination.

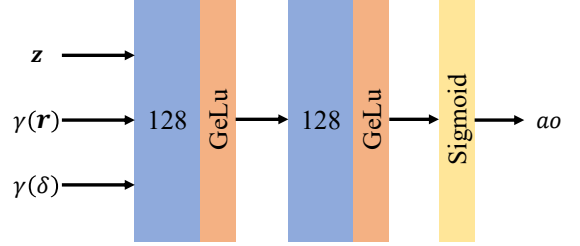


Figure A. **Architecture of the Shadow Decoder.** Our decoder is implemented as a two-layer MLP. Positional information, including distance δ and orientation \mathbf{r} relative to the human, are initially encoded using positional encoding and subsequently passed through a sigmoid function to predict ambient occlusion ao .

A.2. More Details of Light Volume

In training, we use trilinear interpolation to obtain the lighting radiance. However, during validation, the human subject might occupy any position that does not appear during training, which could lead to inaccuracies if we still use trilinear interpolation. To solve it, we track human’s movement during the training stage and record the illuminated light probes using masks. During validation, we employ KNN along with a novel interpolation method [7] to estimate the light radiance. The interpolated weight in Eq. (6) is based on the distance between the Gaussians position and the probe:

$$w(t) = \begin{cases} 2t^3 - 3t^2 + 1 & 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}, \quad (16)$$

where $t = \|\mathbf{x} - \mathbf{p}_i\|/r$, and r represents the maximum distance of the Gaussian \mathbf{x} to the neighboring probes, \mathbf{p}_i is the i th neighbor probes.

A.3. More Details of Loss Functions

The image loss in Eq. (10) consists of human region loss \mathcal{L}_{human} and the entire scene loss \mathcal{L}_{scene} . Both of them are applied by L1 loss, SSIM loss and perceptual loss:

$$\mathcal{L}_{human/scene} = \lambda_{l_1} \mathcal{L}_1 + \lambda_{ssim} \mathcal{L}_{SSIM} + \lambda_{lpips} \mathcal{L}_{LPIS}, \quad (17)$$

where $\lambda_{l_1} = 0.8$, $\lambda_{ssim} = 0.2$, $\lambda_{lpips} = 1.0$. Additionally, the values of λ_h and λ_s in Eq. (10) are set to 1.0 and 10.0.

To enhance the geometry quality of the human, we perform smooth loss on the human mesh in the first stage. We use the laplacian loss and edge loss to regularize the mesh:

$$\mathcal{L}_{mesh} = \lambda_{lap} \mathcal{L}_{lap} + \lambda_{edge} \mathcal{L}_{edge}, \quad (18)$$

where $\lambda_{lap} = 10.0$, $\lambda_{edge} = 1.0$.

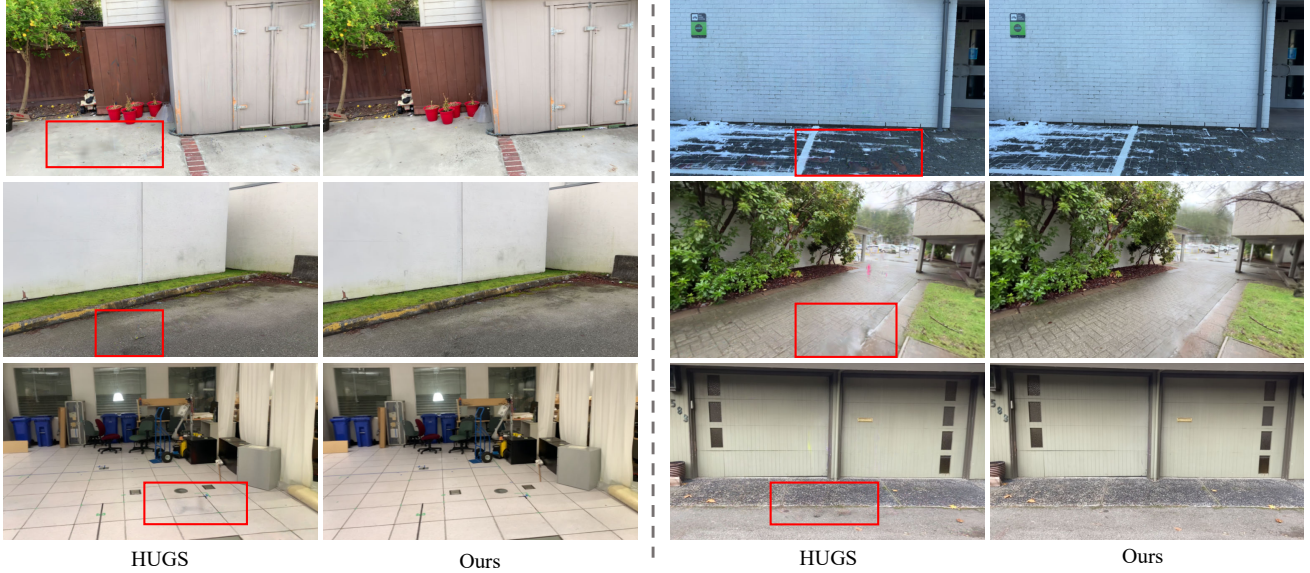


Figure B. Qualitative results comparing our method with HUGS on scene reconstruction.

A.4. Details of Comparable Methods.

1). Human–Scene Reconstruction Methods. 4DGS [11] extends 3DGS into the temporal domain, enabling efficient 4D dynamic scene reconstruction. NeuMan [5] utilizes two separate NeRFs to independently model the static scene and the dynamic human. Vid2Avatar [3] employs signed distance field (SDF) for higher-fidelity human surface reconstruction. HUGS [6] is a 3DGS-based approach that represents humans using tri-planar Gaussians, achieving high-speed reconstruction. The results for Vid2Avatar are provided by its authors, while those for NeuMan and HUGS are rendered using their publicly released models. 2). Human-specific optimization-based methods. Neural Body (NB) [8] trains a set of latent codes anchored to the SMPL mesh and decodes them into color and volume density. HumanNeRF [10] introduces a canonical-volume and motion-field formulation for reconstructing humans from monocular video. Relighting4D [1] models humans as a neural field and infers material properties using pre-learned latent codes, while IntrinsicAvatar [9] performs physically based inverse rendering of clothed humans using explicit Monte-Carlo ray tracing. IRAGA [12] binds 3DGS to a mesh extracted from a precomputed SDF and achieves relighting by optimizing the material properties of the Gaussians.

B. Architecture

We initialize the light volume with $32 \times 32 \times 32$ probes, where each probe represents illumination using spherical harmonics, supplemented by a 48-dimensional lighting feature z . Additionally, a $32 \times 32 \times 32$ mask is employed to indicate the probes are active in training. Fig. A further

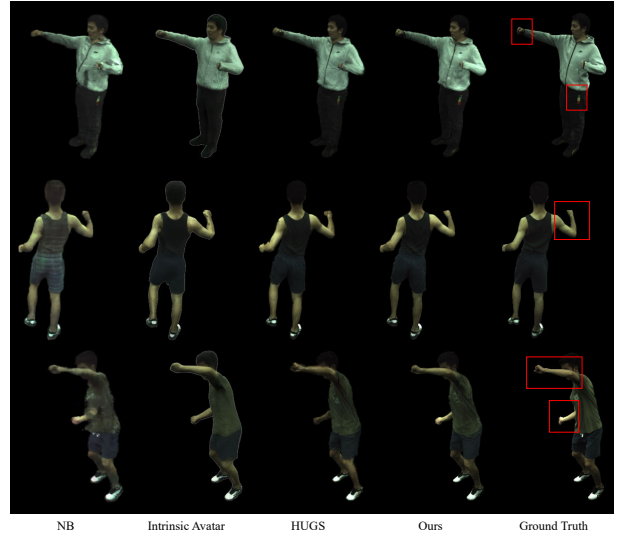


Figure C. Novel view synthesis results produced by our method and baselines on the ZJU-MoCap dataset. This experiment is conducted under a single-view input.

illustrates the architecture of the shadow decoder.

C. More Comparison Results

C.1. Scene Reconstruction

To demonstrate the improved disentanglement of the human and the scene achieved by our method, we compare it with HUGS on the NeuMan dataset for scene reconstruction. The visual comparison results are presented in Fig. B. In the left column of Fig. B, thanks to our scene shadow estimation module, our method demonstrates superior ground

reconstruction, while HUGS, which does not account for human shadows in the scene, inevitably produces blurring and dark artifacts. In the right column of comparisons, our method effectively decouples the human from the scene, whereas HUGS introduces ghosting artifacts of the human within the scene.

C.2. Human Reconstruction

In Sec. 4.2.1, we compare our method against baseline approaches for human reconstruction on the ZJU-MoCap dataset. The quantitative results demonstrate that our method consistently outperforms others across all metrics. To complement these findings, we provide additional qualitative comparisons. As shown in Fig. C, in row 1, our method captures finer high-frequency details. In row 2, our method significantly reduces artifacts in the human arms compared to the baseline methods. Finally, in row 3, our method effectively preserves more accurate lighting information.

C.3. Efficiency of Shadow Estimation.

To further verify the efficiency of our implicit shadow estimation module, we compare the runtime of our method with the ray-tracing-based approach in R3DGS [2], evaluated with 16 and 32 ray samples, respectively. For a fair comparison, the baseline only performs ray tracing on scene Gaussians within the axis-aligned bounding box around the human, which is consistent with our setting. As shown in Tab. A, our approach achieves a speedup of $5\times$ to $10\times$ over the baseline.

Table A. Runtime comparison of our shadow estimation module and the ray-tracing-based shadow approach (RT-16 / RT-32 denotes 16 and 32 ray samples, respectively).

Method	Ours	RT-16	RT-32
Runtime	11ms	59ms	113ms

D. More Ablation study Results

D.1. Ablation studies on Two-stage Strategy

We further conduct ablation studies on the two-stage human reconstruction strategy. As shown in Tab. 4, the two-stage strategy can improve the reconstruction quality. Fig. D shows the normal qualitative comparison: without the two-stage optimization scheme for the human mesh, the estimated surface normals become overly smooth and inaccurate, making it difficult to represent geometric details such as garments.

D.2. Ablation studies on Depth Loss

In Fig. E, we render the depth map using splatting. The results show that removing depth supervision makes the depth

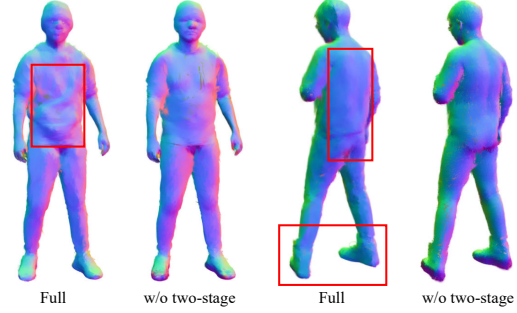


Figure D. **Effect of two-stage strategy.** The two-stage strategy enables the reconstruction of more accurate human geometry.

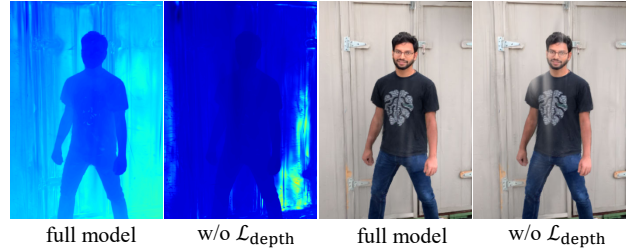


Figure E. **Effect of depth regularization \mathcal{L}_{depth} .** The regularization can further optimize the relative positions between the human and scene.

relationship between the human and the scene ambiguous, leading to confusion in the rendered images. The depth loss helps establish plausible entire scene geometry, which addresses the contact issues caused by the close proximity between the human and the background.

E. More Visualizations

E.1. Transfer Humans to Different Scenes

In Sec. 4.2.2, we show the application of human-scene transfer. Our supplementary materials provide further detailed examples of animated human transfer into various scenes. As shown in Fig. F, our method renders the human with distinct lighting effects at different locations within the scene. For additional results, please refer to our supplementary videos.

E.2. Human Relighting

Fig. G presents human relighting results on NeuMan dataset under different environment maps. Please check the supplementary videos for them.

E.3. Intermediate Outputs Visualization

To emphasize the scene illumination and shadow modeling capabilities of our method, we visualize the intermediate outputs across five sequences from the NeuMan dataset in

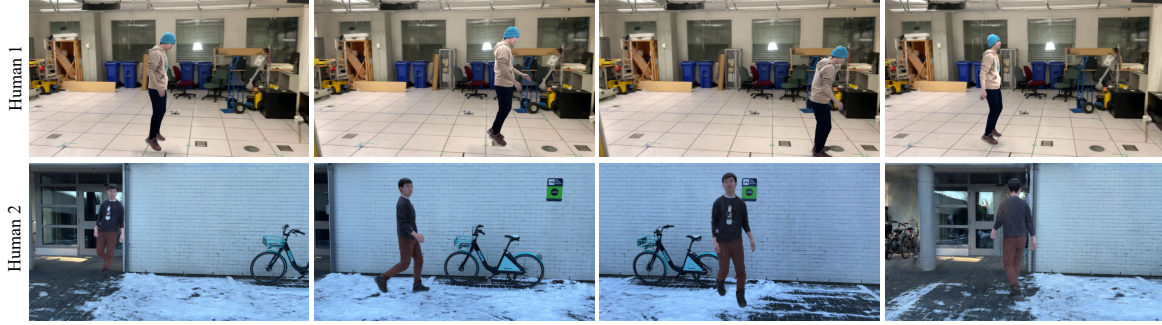


Figure F. Visualization of human in different scenes with novel pose and correspondent lighting condition.



Figure G. **More qualitative results of human relighting.** Our method synthesizes realistic relighting results of different human on the NeuMan dataset under various lighting conditions.

Fig. H. Column 1 and Column 2 present the human rendered directly with the albedo and normal attribute. Column 3 displays the lights rendered with incoming light radiance on human Gaussians and further processed through gamma correction. It shows that our method effectively captures the scene local lighting condition, enabling PBR rendering of humans with diverse appearances in column 4. Furthermore, in column 6 of Fig. H, compared to column 5, more pronounced shadow effects can be observed around the human feet, especially in row 1 and row 3. This illustrates that our scene shadow estimation module can effectively decouple dynamic human shadows from the scene.

In summary, our method effectively disentangles scene lighting, shadows, and human materials while capturing local lighting details. This capability allows our approach to facilitate scene replacement and extend to relighting applications.

References

- [1] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. [2](#), [3](#), [7](#)
- [2] Jian Gao, Chun Gu, Youtian Lin, Zhihao Li, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussians: Realistic point cloud relighting with brdf decomposition and ray tracing. In *European Conference on Computer Vision*, pages 73–89. Springer, 2024. [2](#), [3](#), [4](#), [5](#)
- [3] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. [1](#), [2](#), [6](#)
- [4] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20418–20431, 2024. [1](#), [2](#), [3](#)
- [5] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel,

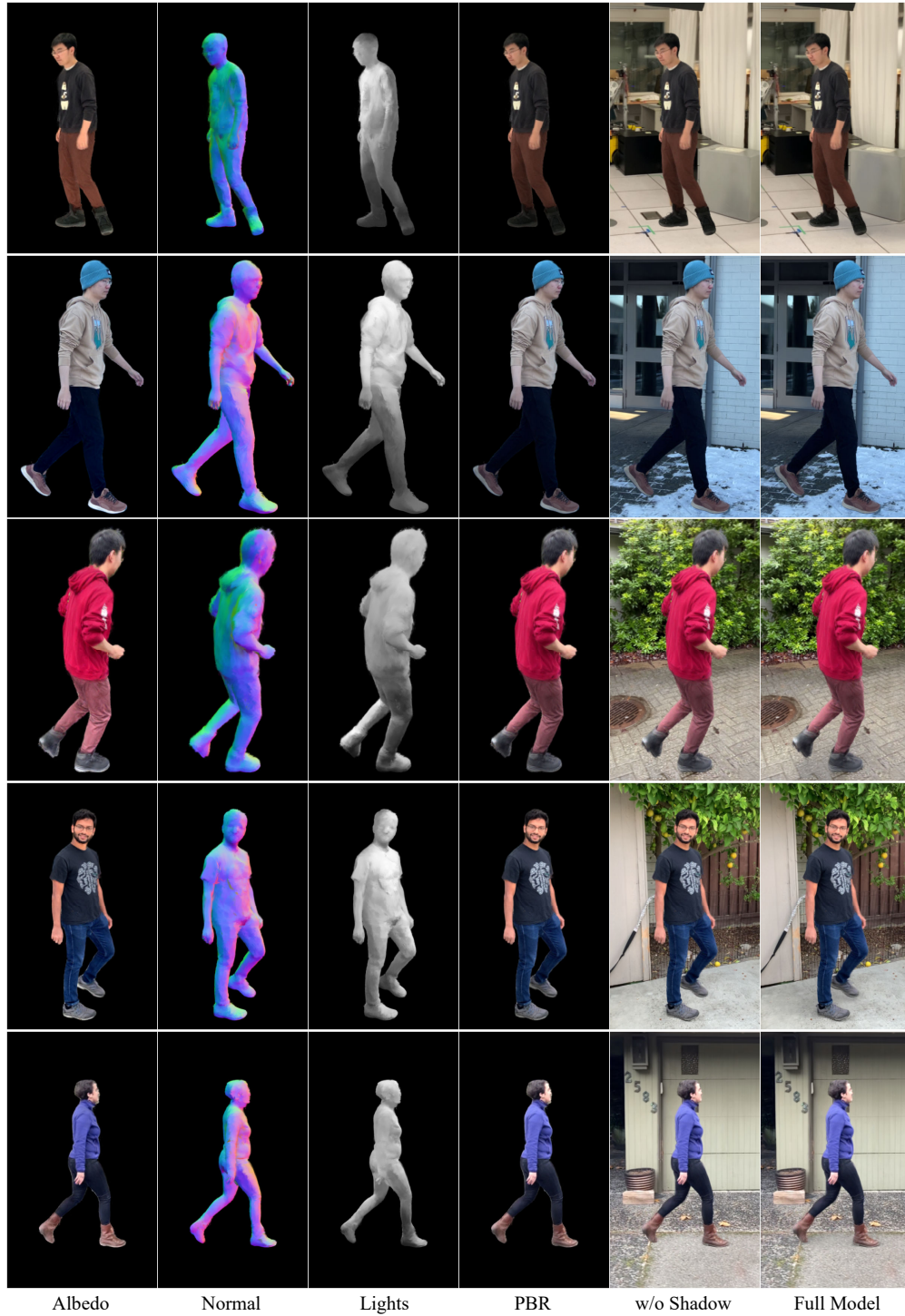


Figure H. Visualization of the intermediate outputs in our framework.

and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. [1](#), [2](#), [5](#), [6](#)

[6] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian

splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [1](#), [2](#), [5](#), [6](#)

[7] Jaakko Lehtinen, Matthias Zwicker, Emmanuel Turquin, Janne Kontkanen, Frédo Durand, François X Sillion, and Timo Aila. A meshless hierarchical representation for light

- transport. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. [1](#)
- [8] 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](#)
 - [9] Shaofei Wang, Bozidar Antic, Andreas Geiger, and Siyu Tang. Intrinsicavatar: Physically based inverse rendering of dynamic humans from monocular videos via explicit ray tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1877–1888, 2024. [3](#), [7](#), [2](#)
 - [10] 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. [2](#)
 - [11] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. 2024. [1](#), [6](#), [2](#)
 - [12] Youyi Zhan, Tianjia Shao, He Wang, Yin Yang, and Kun Zhou. Interactive rendering of relightable and animatable gaussian avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2025. [2](#), [3](#), [5](#), [7](#)