Generalizable Human Gaussians for Sparse View Synthesis – *Supplementary Material* –

Youngjoong Kwon[1], Baole Fang[1*], Yixing Lu[1*], Haoye Dong[1], Cheng Zhang[1], Francisco Vicente Carrasco[1], Albert Mosella-Montoro[1], Jianjin Xu[1], Shingo Takagi[2], Daeil Kim[2], Aayush Prakash[2], and Fernando De la Torre[1]

[1]Carnegie Mellon University [2]Meta Reality Labs

https://humansensinglab.github.io/Generalizable-Human-Gaussians/

A Appendix - Overview

This appendix is organized as follows: Sec. B discusses the limitations and future works; Sec. C presents the societal impacts our work can have; Sec. D shows additional results including video results, comparison with single-view methods, ablations on number of outer scaffolds, ablation study with different loss supervision, ablations on the number of input views at inference time, and runtime at inference. Sec. E provides information regarding reproducibility, which includes implementation details.

B Limitations and Future Works

Although our method achieves state-of-the-art results in terms of visual quality and runtime, it is not free from limitations. (1) While our method effectively compensates for minor inaccuracies in SMPL-X estimations through the use of multi-scaffolds, significant deviations in SMPL-X from the input images could compromise the quality of our results, as our Gaussians are anchored to the SMPL-X surface. (2) Currently, the number of scaffolds is determined empirically. It would be an interesting direction to explore adaptive scaffolds based on subject attributes (e.g., loose or tight clothing). (3) The performance of our inpainting network is constrained by the small number of ground truth texture maps available during training, which in turn limits its ability to generate detailed hallucinations when given a single-view input. Therefore, integrating and fine-tuning generative models trained on extensive datasets (e.g., Stable Diffusion model [7]) could substantially improve our network's hallucination capabilities and generalizability, which is a promising direction for future work.

* Equal contribution

C Societal Impacts

Our proposed method can push immersive entertainment and communication to a more affordable setting. For example, our work has the potential to enhance the accessibility of telepresence experiences by facilitating the creation of avatars from minimal RGB images. Moreover, the technology presents benefits to film and game production by enabling efficient synthesis of large-scale 3D human avatars with low costs.

However, our work might also introduce potential challenges, primarily related to the accessible creation of realistic human images. This could lead to deep-fake human avatars on social media, with implications for misinformation and the degradation of trust in digital content. To mitigate such risks, it is urgent to promote ethical guidelines and regulations on synthetic media. We strongly appeal transparent use of such technology as it should align with societal interests and foster trust rather than skepticism.

D Additional results

D.1 Video results

Video results of comparison with the state-of-the-art baselines on the in-domain generalization task (i.e., trained and tested on THuman 2.0 dataset [10]) and cross-dataset generalization task (i.e., trained on THuman 2.0 and tested on RenderPeople [6]) can be found in the project website[*]. For the in-domain generalization task, we compare our GHG with (1) human template-conditioned NeRF, generalization from sparse view methods NHP [3] and NIA [4], and (2) generalizable 3D Gaussian Splatting for human rendering method GPS-Gaussian [11]. Note that GPS-Gaussian is trained and tested with 5 input views due to the rectification requirement. NHP, NIA, and ours are trained and tested with 3 input views. For the cross-dataset generalization task, we show comparison with our main baselines NHP and NIA. Our method can recover sharp and fine details compared to human template-conditioned NeRF baselines. Due to the lack of full 3D prior, GPS-Gaussian suffers in maintaining multi-view consistency between the novel views generated using different input views. On the other hand, ours maintains robust and accurate geometry reconstruction utilizing the 3D human template.

D.2 Comparison with single-view methods

Fig. 1 shows comparisons with SOTA single-view reconstruction methods that are based on 3D human prior: ECON [8], TECH [2], and SiTH [1]. We used their officially released implementation for the comparison. Our sparse-view work outperforms in terms of accuracy and faithfulness to the observed data, as can be

[*] https://humansensinglab.github.io/Generalizable-Human-Gaussians

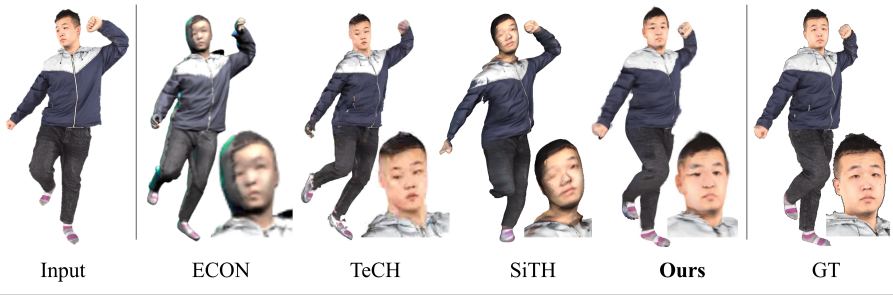

Fig. 1: Comparison with single-view reconstruction methods: ECON [8], TeCH [2], and SiTH [1]. Our method outperforms the baselines in terms of faithfulness to the given observation.

Table 1: Ablation study on the number of outer scaffolds used. We trained and tested variants with different numbers of scaffolds that are outside the original SMPL-X surface. The variant with only the base template is denoted as "0 scaffold". The performance increase is saturated as more than 5 outer scaffolds are used.

# Out scaffolds.	PSNR↑	LPIPS↓	FID↓
0	22.30	145.74	84.38
1	**22.77**	139.16	75.66
2	22.28	137.65	73.54
3	21.87	136.38	65.19
4 **(Ours full)**	21.90	**133.41**	**61.67**
5	22.13	134.73	63.80
6	22.09	135.52	64.81

seen in Fig. 1. Also, the single-view methods either require per-subject optimization (ECON, TeCH) or run at relatively slow speed (e.g., ECON 3 min / TeCH 4 hr / SiTH 2 min). On the other hand, ours is a feed-forward method that runs at $4fps$, which is $\times 480$ faster than SiTH.

D.3 Ablations

Ablation on the number of scaffolds. In Tab. 1, we study the impact of number of outer scaffolds. Variants with different number of outer scaffolds are trained and tested. The performance increase is saturated as more than 5 outer scaffolds are used. Therefore, we use 4 outer scaffolds as our final model. In Fig. 2, we show how the number of scaffolds affects the reconstruction of offset details such as hair (a,c) and loose clothing (b,c).

Ablation on the supervision. Tab. 2 shows the impact of different loss supervision employed during training. Note that our variant with L_1-only supervision (Tab. 2-a) already outperforms the human template-conditioned generalizable NeRF methods NHP and NIA, which are also trained with L_1-only supervision,

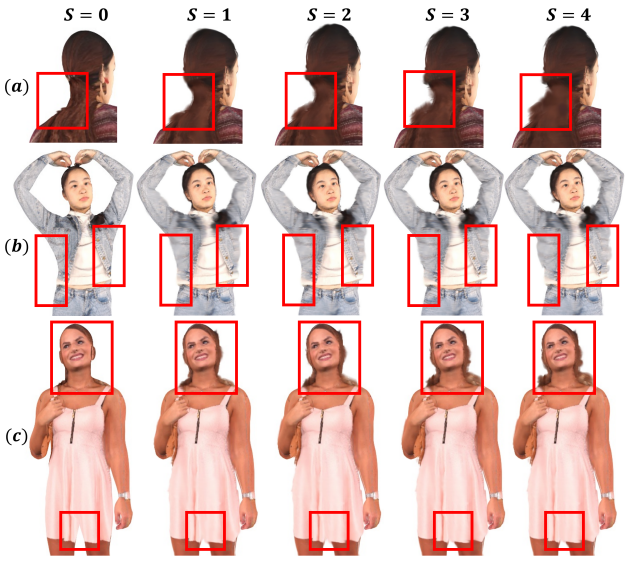

Fig. 2: Multi-scaffold helps reconstruct hair and loose clothing. S denotes the number of outer scaffolds.

in terms of perceptual metrics LPIPS and FID. This validates that our gain is not only from the different supervision but also from our proposed multi-scaffold. Our full model that leverages multi-view supervision with L_1, SSIM, and mask loss achieves the highest performance on the perception-based metrics. Note that multi-view supervision is possible by leveraging the fast 3D Gaussian splatting.

Ablation on the number of input views at inference. We trained our model using 3 input views and tested with different number of input views at inference time in Tab. 3. The performance improves as more observations are available. However, note that our performance when only given two views is still comparable to the 3-view results. This demonstrates the effectiveness of our method under sparse view setting.

Performance on the randomly selected input views. During evaluation, we followed the convention of previous sparse view 3D human reconstruction works [3, 4] that use 3 uniformly distributed inputs. However, we additionally ran the evaluations given 3 random views 10 times and computed the mean metrics. We verified that the performance difference between the uniformly and randomly sampled inputs is minimal – PSNR is 1.5%, and LPIPS is 0.3%.

D.4 Runtime at inference

Our GHG runs at $4fps$ for rendering a single $1K$ (1024×1024) image on a single NVIDIA RTX A4500 GPU. However, note that inpainting network takes most

Table 2: Ablation study on the supervision. ✗/✓indicates completely remove/keep the loss supervision. Our L_1-only supervision result (a) still outperforms the human template-conditioned NeRF methods NHP and NIA, which are also trained with L_1-only supervision. This validates the effectiveness of our proposed multi-scaffold.

	L_1	SSIM	Mask	Multi-view	PSNR↑	LPIPS↓	FID↓
NHP	✓	✗	✗	✗	**23.32**	184.69	136.56
NIA	✓	✗	✗	✗	23.20	181.82	127.30
a	✓	✗	✗	✗	23.05	142.57	71.97
b	✓	✓	✗	✗	22.69	136.44	69.50
c	✓	✓	✗	✓	22.03	134.82	62.04
Ours full	✓	✓	✓	✓	21.90	**133.41**	**61.67**

Table 3: Ablation study on the number of input views at inference. We trained our model using 3 input views, and tested with different numbers of input views at inference time. The performance improves as more observations are available.

# Inputs.	PSNR↑	LPIPS↓	FID↓
1	20.08	152.54	99.13
2	21.79	132.61	78.56
3	21.90	133.41	61.67
4	22.01	133.68	53.40
5	**22.07**	**131.80**	**35.00**

of our runtime (74%). Without the inpainting network, ours runs at $15fps$. More efficient inpainting model can be explored to further reduce the runtime.

The detailed breakdown of runtime is as follows. Our pipeline can be divided into three stages: (1) constructing multi-scaffold (2) Gaussian parameter map generation (3) rasterization. **(1) Constructing multi-scaffold:** RGB map for each scaffold is aggregated on the UV space of human template. Our inpainting network inpaints the missing regions of the innermost scaffold RGB map in 180.89 ms. **(2) Gaussian parameter map generation:** Multi-Gaussian parameter maps are generated in 57.97 ms. **(3) Rasterization:** Rasterization takes 5.78 ms. In total, GHG takes 244.65 ms to render a single $1K$ image.

We would like to highlight that our method runs faster than the sparse-view generalizabl human NeRF methods NHP and NIA ($0.01fps$ to render a single 1K image) while outperforming their visual quality.

E Implementation details

E.1 Gaussian parameter map generation

The architecture design of our Gaussian parameter map generation network is presented in Fig. 3. Our network is composed of two encoders $\mathcal{E}_{\text{appr}}$, $\mathcal{E}_{\text{geo}}$ and one decoder $\mathcal{D}_{\text{dec}}$. The feature maps extracted by $\mathcal{E}_{\text{appr}}$ and $\mathcal{E}_{\text{geo}}$ are added together before being fed into $\mathcal{D}_{\text{dec}}$. Moreover, $M_{\mathbf{s}}$ and M_{α} are sent into *Softplus* and

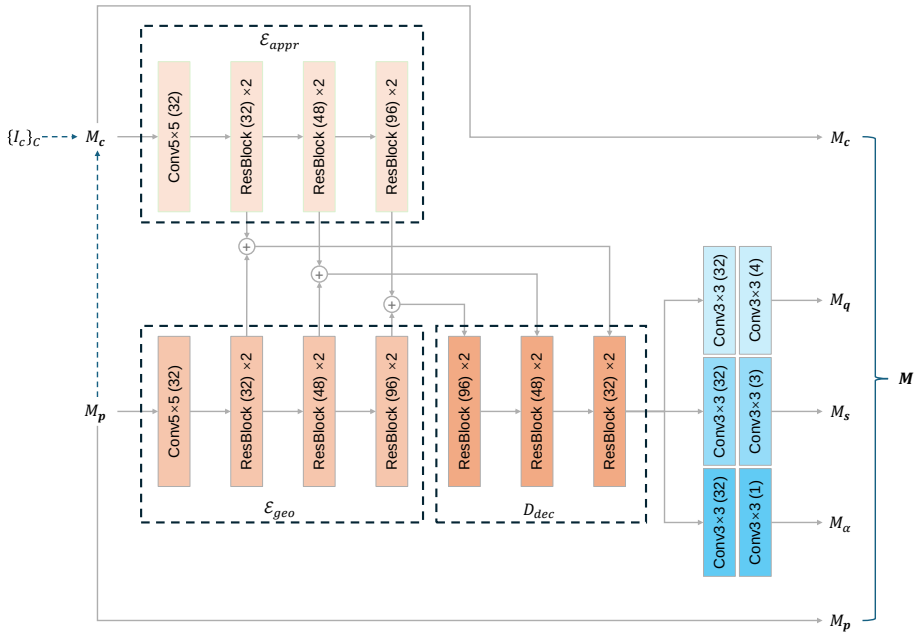

Fig. 3: Network architecture for Gaussian parameter map generation.

Sigmoid activation layers, respectively, after the convolution layers. Note that in the figure, the number following each layer name and sitting in the bracket denotes its output channel size.

E.2 Inpainting

Pseudo ground truth generation To create the pseudo ground truth texture map on the SMPL-X UV space, we follow the approach proposed in Lazova et al [5]. The process is illustrated in Fig. 4. For each point on the SMPL-X model, we identify the nearest point on the scanned object. Next, we determine the corresponding position of this point on the scan's UV map. We then transfer the color from this position on the scan's UV map to the corresponding location on the SMPL-X's UV map.

Network architecture Fig. 5 shows the inpainting module architecture. The inpainting network follows the DeepFillv2 design [9]. The inpainting network is composed of a generator $\mathcal{G}_{\text{inpaint}}$ and a discriminator $\mathcal{D}_{\text{inpaint}}$. In the generator, all convolutions are gated convolutions with a kernel size of 3×3 if not specified, where *GatedConv*, *DilateGatedConv*, *GatedConvDown*, *GatedConvUp* have a stride of 1, 1, 2, 0.5, respectively. The four *DilateGatedConv* layers in *DilatedBlock* have a dilation of 2, 4, 8, 16, respectively. The *Attention* layer is a self-attention layer. In the discriminator, all convolutions are common 2D convolutions, where *Conv*, *ConvDown* have a stride of 1, 2, respectively. Besides, all

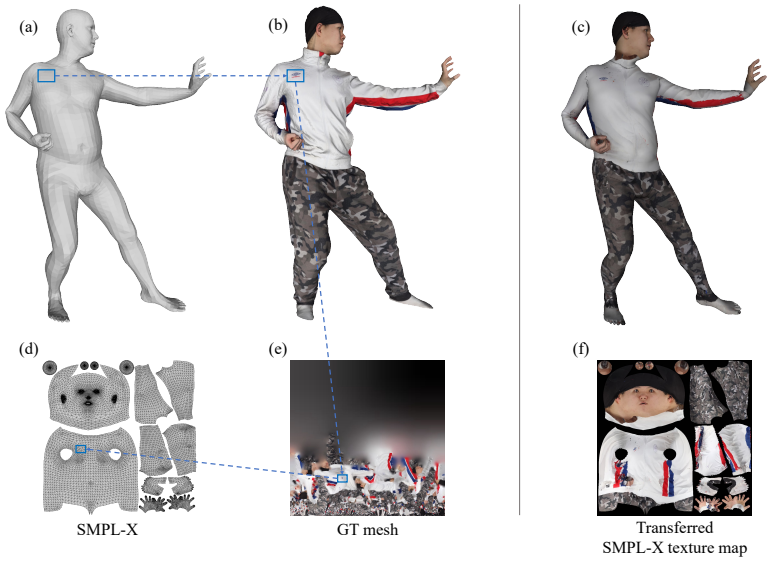

Fig. 4: Illustration of texture transfer on to the SMPL-X UV space. For each point on the SMPL-X model (a), the nearest point on the scanned mesh (b) is found. Then, we get the corresponding position of this point on the scan's UV map (e), which will be mapped to the matching location on the SMPL-X's UV map (d). Resulting on the transferred texture map (f) and the colored mesh (c).

convolution layers are followed by ELU activation. Note that in the figure, the number following each layer name and sitting in the bracket denotes its output channel size.

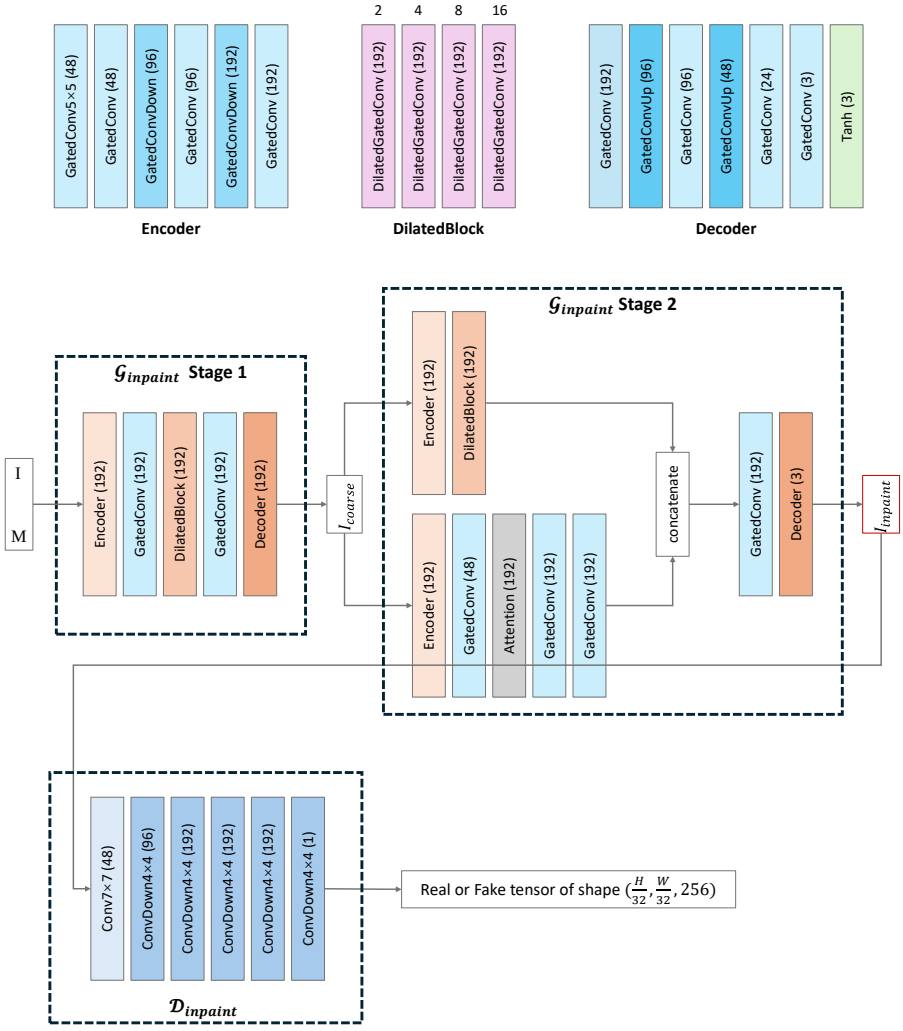

Fig. 5: Inpainting network.

References

1. Ho, H.I., Song, J., Hilliges, O.: 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, 3
2. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: Tech: Text-guided reconstruction of lifelike clothed humans. arXiv preprint arXiv:2308.08545 (2023) 2, 3
3. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. Advances in Neural Information Processing Systems **34**, 24741–24752 (2021) 2, 4
4. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural image-based avatars: Generalizable radiance fields for human avatar modeling. In: International Conference on Learning Representations (2023) 2, 4
5. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: 2019 International Conference on 3D Vision (3DV). pp. 643–653. IEEE (2019) 6
6. RenderPeople. http://renderpeople.com (2018) 2
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1
8. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: Econ: Explicit clothed humans optimized via normal integration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 512–523 (2023) 2, 3
9. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. arXiv preprint arXiv:1806.03589 (2018) 6
10. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021) 2
11. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 2