Generalizable Human Gaussian Splatting via Multi-view Semantic Consistency

Supplementary Material

1. Introduction

In this supplementary material, we provide additional explanations of the proposed method. First, details of implementation are described in Section 2. In the following, the detailed architecture of the proposed method is demonstrated in Section 3. Further ablation study on the proposed method is presented in Section 4. Finally, additional qualitative results with discussion are given in Section 5.

2. Implementation Details

Pre-trained backbone configuration. To encode multi-view inputs into latent embeddings, we adopt the pre-trained VGGT encoder [7], which consists of 24 transformer layers. As follows [7], latent embeddings are taken from the 4^{th}, 11^{th}, 17^{th}, and 23^{rd} layers. Each layer produces latent embeddings $E \in \mathbb{R}^{V\times(H/p)\times(W/p)\times C}$ (where V is the number of views, $p = 14$ is the patch size, $H = W = 518$, and $C = 2048$). In addition, we extract semantic features from the last layer of the DINO vision transformer [4], which is included in the VGGT encoder.

Details of unprojection process. To associate each latent embedding with a 3D position, we first downsample the depth map estimated from each viewpoint to the spatial resolution of the latent embeddings (i.e., $(H/p) \times (W/p)$) by using nearest-neighbor interpolation. The camera intrinsic parameters are scaled accordingly, where both the focal length and the principal point are divided by the same downsampling factor. By using rescaled results of the depth map and the camera intrinsic parameters, each pixel is unprojected into a 3D point in the camera coordinate system. The points are subsequently converted into the world coordinate system by leveraging the camera extrinsic parameters. Since each latent embedding corresponds to the same spatial index on the downsampled depth map, its 3D position is determined by the unprojected result.

3. Architecture Details

Recalibration of latent embeddings. We provide additional details to recalibrate latent embedding via cross-view attention in the 3D space. The proposed method follows the structure of a standard transformer encoder [2], while being extended to account for the spatial relationship and the semantic consistency between latent embeddings. For each embedding, we identify its local neighborhood using K-nearest neighbor (KNN) search with $K = 64$. The latent embedding and its neighbors are then normalized and independently projected through linear layers to obtain query, key, and value representations. The cross-view attention scheme is applied to each local neighborhood to generate recalibrated latent embeddings (the detailed explanation is

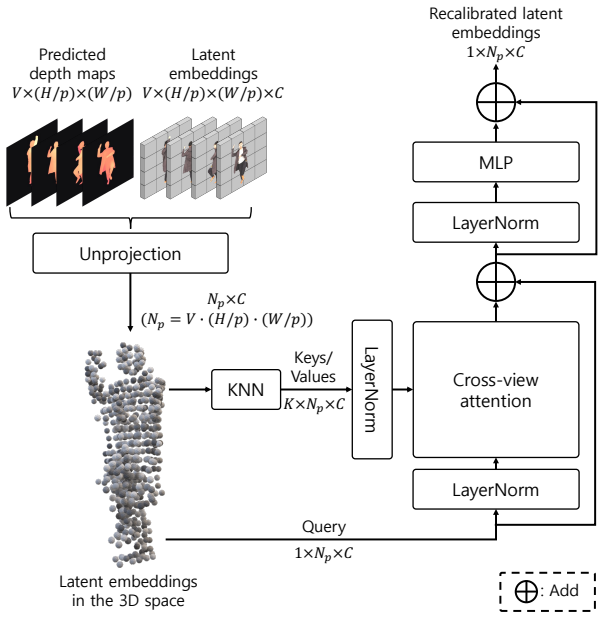

Figure 1. The detailed architecture for recalibration of latent embedding via cross-view attention in the proposed method.

Attribute	Layer Type	Input Dim.	Output Dim.
Position offset	1×1 Conv.	32	128
	1×1 Conv.	128	32
	1×1 Conv.	32	3
Scale	1×1 Conv.	32	128
	1×1 Conv.	128	32
	1×1 Conv.	32	3
	Softplus	3	3
Rotation	1×1 Conv.	32	128
	1×1 Conv.	128	32
	1×1 Conv.	32	4
	Norm.	4	4
Opacity	1×1 Conv.	32	128
	1×1 Conv.	128	32
	1×1 Conv.	32	1
	Sigmoid	1	1

Table 1. The detailed architecture of Gaussian header for each attribute.

Figure 2. Results of novel view synthesis via generalizable human Gaussian splatting on the THuman2.0 [9] dataset. (a) Input images. (b) Results by GPS-Gaussian [10]. (c) Results by GHG [3]. (d) Results by RoGSplat [8]. (e) Results by the proposed method. (f) Ground truth.

provided in Section 3.3 of the manuscript). The output of our cross-view attention is processed through a feed-forward network with residual connections. The detailed architecture is illustrated in Fig. 1.

Gaussian regression. After obtaining the recalibrated latent embedding, these embeddings are projected back into the 2D grid and processed by a DPT-style decoder [6] to generate per-pixel Gaussian attributes. The decoder produces a dense feature map $X_d \in \mathbb{R}^{H\times W\times D}$ (where $D = 32$), from which each Gaussian attribute is predicted using an independent regression head. Each head is implemented as a sequence of 1×1 convolutions with ReLU activations, followed by a final projection layer whose output dimension matches the corresponding Gaussian attribute. The detailed architecture for each header is provided in Table 1.

4. Ablation Study

In this Section, we analyze the influence of adjusting the number of neighbors (i.e., K in KNN search) for computing cross-view attention of latent embeddings. For this experiment, three values of $K \in \{32, 64, 128\}$ are tested on the THuman2.0 [9] dataset with four-view inputs. As shown in Table 2, the configuration with $K = 64$ achieves the highest overall performance. Increasing K from 32 to 64 yields the meaningful improvement in PSNR, whereas using $K = 128$ does not provide additional gains. Based on this observation, we adopt $K = 64$ as the default value in all experiments.

# Neighbors (K)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
$K = 32$	30.57	0.9700	0.0344
$K = 64$	30.93	**0.9710**	**0.0334**
$K = 128$	**31.01**	0.9708	0.0335

Table 2. Performance analysis of the proposed method according to the number of neighbors (K) used in KNN search.

5. More Results and Limitations

5.1. Qualitative Results

To show the effectiveness of the proposed method, additional qualitative comparisons both on THuman2.0 [9] and

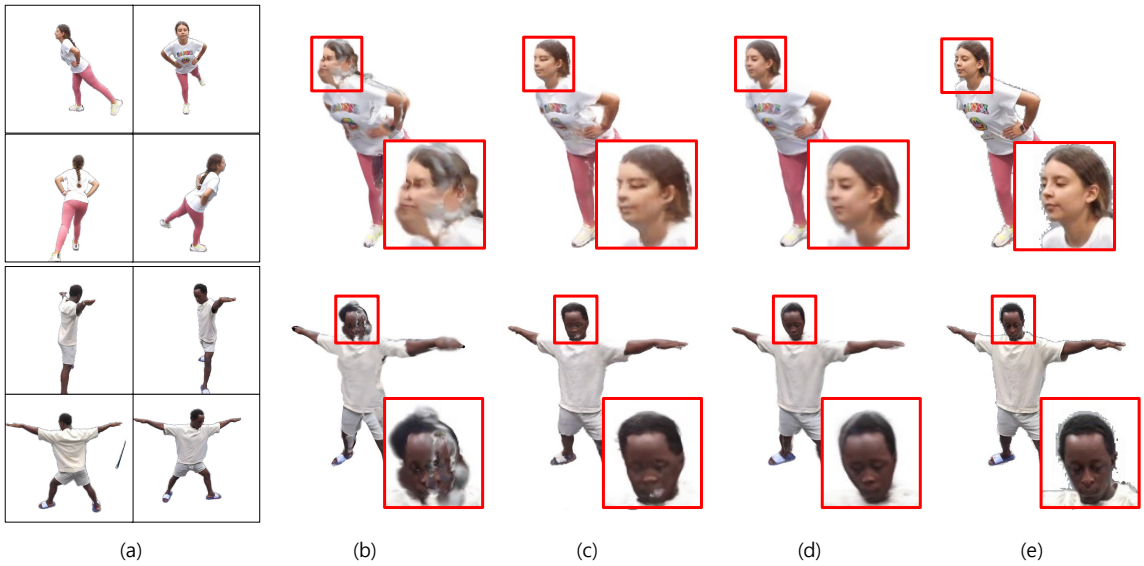

Figure 3. Results of novel view synthesis via generalizable human Gaussian splatting on the HuMMan [1] dataset. (a) Input images. (b) Results by GHG [3]. (c) Results by RoGSplat [8]. (d) Results by the proposed method. (e) Ground truth.

HuMMan [1] datasets are presented in Figs. 2 and 3, respectively. Specifically, the proposed method successfully represents textural details without significant distortions in complicated regions, e.g., facial regions, compared to previous methods (see the 1st rows in Figs. 2 and 3). In addition, the proposed method accurately renders boundaries of loose clothing, where existing methods produce incomplete reconstruction (see the 3rd row in Fig. 2).

5.2. Limitations and Future Work

Although the proposed method reliably improves the quality of human rendering, it may still struggle to reconstruct very thin or highly articulated structures, such as fingers. As shown in Fig. 4, these regions appear slightly blurred in the rendered results. This limitation arises because thin structures occupy only a small number of valid pixels across different input views, which limits the influence to the computation process of cross-view attention and thus reduces the reliability of the semantic consistency. As a result, the recalibrated latent embedding for such regions may lack sufficient support to precisely regress Gaussian attributes, thus leading to loss of details in the rendering result. Extension of the proposed method to handle dynamic inputs or editable representations remains as the future work.

6. Ethics Statement

We use publicly available human datasets, including THuman2.0 [9], ZJU-MoCap [5], and HuMMan [1], all of which provide properly released data for research purposes. Our work focuses on generalizable human Gaussian splatting from sparse-view inputs, a technology that can benefit applications in graphics, telepresence, and virtual environments. However, as this technique enables the reconstruction of a realistic human appearance from limited visual observations, it may raise concerns related to privacy, identity misuse, or unintended replication of personal likeness. We emphasize that our method is intended solely for research use and should not be applied to scenarios involving unauthorized digital reproduction of individuals. Responsible deployment and adherence to ethical guidelines are strongly encouraged.

Figure 4. Limitations of the proposed method. Note that thin structures, e.g., fingers, remain challenging to reconstruct, which leads to slightly blurred results.

References

[1] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *Proc. Eur. Conf. Comput. Vis.*, pages 557–577, 2022. 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent.*, pages 1–22. 1

[3] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable human gaussians for sparse view synthesis. In *Proc. Eur. Conf. Comput. Vis.*, pages 451–468, 2024. 2, 3

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1

[5] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9054–9063, 2021. 3

[6] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. Int. Conf. Comput. Vis.*, pages 12179–12188, 2021. 2

[7] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5294–5306, 2025. 1

[8] Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. RoGSplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5980–5990, 2025. 2, 3

[9] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5746–5756, 2021. 2, 3

[10] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19680–19690, 2024. 2