Learning Efficient and Generalizable Human Representation with Human Gaussian Model

Supplementary Material

6. Implementation Details

Fine-tuning multi-view diffusion. We utilize the Wonder3D [23] as the multi-view diffusion model. As the Wonder3D model is designed for objects and trained on Objaverse [7], it has limited knowledge about humans and directly applying it into our architecture would lead to dissatisfactory results. To mitigate this problem, we fine-tuned Wonder3D with the MvHumanNet dataset. To ensure training efficiency, we only selected 600 humans and used 30 poses for each scan, summing to 18K training data. We input the front view and supervise the output with corresponding front, back, left and right views. The resolution is set to 256, and we use a learning rate of 5e-5 at the mixed training stage and 2.5e-5 at joint training stage. The training converges on 8 $\times$ Nvidia A800 GPUs in 9 days, with a batch size of 4 per GPU. Though improved, the overall result is still non-optimal, leading to the gap between the monocular setting and the multiview setting in our experiments.

Fine-tuning Gaussian reconstruction model. To obtain an initial set of Gaussians mentioned in Section 3.2:

$$G^t = \{g_m^t\}_{m=1}^M, \quad g_m^t = \{\mu_m^t, f_m^t\}, \tag{19}$$

we leverage a fine-tuned LGM model. We use the same dataset mentioned in Section 4.1, and fine-tuned the "large"" LGM with an input resolution of 256 and an output Gaussian resolution of 128 per view. We follow the original LGM configuration for the learning rate and batch size. The training converges on 8 $\times$ Nvidia A800 GPUs in 5 days.

Training details. We train our HGG modules with the fine-tuned LGM model frozen. We uniformly sample 8 frames from each video and use the 8 frames as input for our module. The learning rate is set to 4e-4, gradient clip is 1.0, batch size to 1 per GPU and gradient accumulation steps to 8. We trained our model on 8 $\times$ Nvidia A800 GPUs, and it converges in 18 hours.

Evaluation split. Our evaluation split is separated from the training split. We randomly selected 10 scans in the MvHumanNet as the evaluation split. Their IDs are listed as follows: 200102, 200114, 200125, 200134, 200137, 200151, 200535, 202148, 202209, 204157.

7. More Analysis

7.1. Efficiency

Though processing large amount of information across the frames, HGG is highly-efficient thanks to the design of learnable queries, negligible compared with the LGM U-net. In this paragraph, the efficiency of each module will be theoretically analyzed.

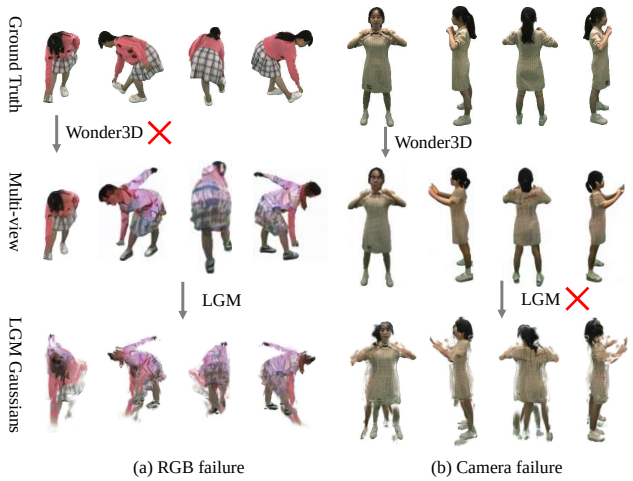

Figure 5. Fail cases. (a) Wonder3D fails to generate reasonable back and side views, resulting in the failure for LGM Gaussian reconstruction. (b) Wonder3D generates multi-views of good quality, but LGM reconstruction fails due to inconsistent camera constraints provided by Wonder3D.

The intra-node transformer enables efficient communication with the Gaussians. If attention is directly applied to the union of all Gaussian sets, the complexity would be $O(M^2T^2D^2)$, M is the number of feed-forward Gaussians per frame, T is the number of the frames, and D denotes the dimension of the features. Given that N is often at the 10,000 level and a video typically have hundreds of frames, it would become an unacceptable computation bottleneck both for time and memory. With our HGG, Gaussians are collected by the mesh vertices, and the complexity becomes linear when each mesh node applies cross attention with the affiliated Gaussians with learnable queries:

$$O((MT/N) * N * D^2) = O(MTD^2) \tag{20}$$

This would be over 10^6 times more efficient than vanilla cross attention between Gaussians.

On the other hand, the computation complexity of the inter-node attention is rather small, thanks to the connectivity given by mesh. Empirically, a node has approximately 10 neigbours, so the complexity is only $O(10ND^2)$, which is negligible compared with other modules.

Figure 6. More qualitative results on novel pose animation. The human avatar in various poses indicates the high-quality of our reconstructed 3D avatar.

Table 3. Quantitative results of 50 testing examples.

	Deepfashion			MvHumanNet		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
LGM	18.01	0.846	0.196	19.54	0.887	0.129
Ours (LGM)	20.31	0.901	0.172	21.82	0.892	0.112
IDOL	20.24	0.904	0.174	21.03	0.894	0.116
Ours (IDOL)	**22.38**	**0.912**	**0.154**	**23.59**	**0.930**	**0.092**

7.2. Monocular Setting

As introduced in the limitations, we observe a gap of 3.4 dB in PSNR between monocular settings and multi-view settings. Though we achieved SOTA in monocular setting, the quality is still far from downstream applications. As illustrated in Figure 5, we found that the fail cases largely derives from (1) the total failure of Wonder3D to generate novel views. (2) generated images do not follow camera constraint strictly, causing misalignments across views. These two reasons account for the failure in Gaussian initialization with LGM, and thus lead to corrupted results.

We attribute this issue to a lack of open-sourced real-world human diffusion models. Such work will largely fuel the field of single-view human reconstruction.

7.3. Comparison with new methods

IDOL can replace LGM to serve as the single-frame reconstruction module in our pipeline. Therefore, we further evaluate our method with this module. As shown in Table 3, our method surpasses the SOTA method and achieves better results with the stronger backbone IDOL. Yet, AniGS has not open-sourced their codes or test splits.

We conduct experiments on Deepfashion, an in-the-wild fashion clothing dataset. As shown in Table 3 and Figure 7, our model achieves consistent performance improvements with different reconstruction modules.

8. More Visualization Results

We present more visualization results for both novel view synthesis and novel pose animation in Figure 6 and 8.

9. Broader Impacts

Our model's capacity to generate high-quality 3D animatable avatars raises substantial privacy risks. To address

(a) Deepfashion (b) MvHumanNet

Figure 7. Visualization on Deepfashion and MVHumanNet.

these, the creation of ethical guidelines and legal frameworks is imperative. This necessitates close collaboration among researchers, developers, and policymakers. Researchers should embed ethical considerations in development, while developers must implement privacy-centric practices. Policymakers need to craft regulations that define proper use, penalize misuse, and safeguard user privacy. Such collaboration is crucial for promoting the responsible application of this technology.

Figure 8. More qualitative results on novel view synthesis. The novel views in multiple directions indicate the high-quality and potential downstream applications of our reconstructed 3D avatar.