MoGA: 3D Generative Avatar Prior for Monocular Gaussian Avatar Reconstruction

Zijian Dong[1,4*] Longteng Duan[1*] Jie Song[3] Michael J. Black[4] Andreas Geiger[2]
[1]ETH Zürich, Department of Computer Science [2]University of Tübingen, Tübingen AI Center
[3]HKUST(GZ)&HKUST [4]Max Planck Institute for Intelligent Systems, Tübingen

1. Implementation

1.1. Technical Details

1.1.1. 2D Gaussian Splatting

2D Gaussian Splatting is a technique that represents 3D scenes as a collection of Gaussian distributions. These Gaussians can be efficiently rendered by projecting them onto the 2D image plane. Each Gaussian primitive $\mathcal{G}_k$ is parameterized by five attributes: an opacity $\sigma_k \in \mathbb{R}$, a Gaussian center $\mu_k \in \mathbb{R}^3$, a view-dependent color c_k parameterized by spherical harmonics, in our model we directly employ an RGB color $c_k \in \mathbb{R}^3$ for simplicity, a scale vector $s_k \in \mathbb{R}^2$ for 2D Gaussian Splatting, and a rotation matrix represented by the axis angle vector $r_k \in \mathbb{R}^3$.

For each pixel $\mathbf{x} = (x, y)$, the pixel color is obtained by alpha blending, which is blending the N projected 2D Gaussians within that pixel.:

$$c(\mathbf{x}) = \sum_{i=1}^{N} c_i \mathcal{G}_i(\mathbf{x}) \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j \mathcal{G}_j(\mathbf{x})) \tag{1}$$

where c_i is the color of the i-th projected 2D Gaussian primitive sorted by depth. To render normal maps, we replace the color c_i with the normal of the Gaussian primitives. σ_i represents the opacity values. $\mathcal{G}(\mathbf{x})$ is the evaluated 2D Gaussian value. More details of the evaluation of $\mathcal{G}(\mathbf{x})$ can be seen in [8].

1.1.2. Gaussian Primitives Initialization

We initialize the center of Gaussian primitives at the center of the faces of a densified SMPL-X mesh. The initial rotation of a Gaussian primitive is its tangent plane coordinate. The initial scale of a Gaussian primitive is its average distance to its k nearest neighbors.

1.1.3. Decoder

Our decoder consists of a CNN upsampler and two lightweight CNN decoders. The CNN upsampler upsamples the latent code twice by a factor of $2\times$, each time $X_i = bilinear\_interpolation(conv(conv(X_i)))$, and there is one final convolutional layer. After we have the upsampled latent code from $64 \times 64 \times 32$ to $256 \times 256 \times 32$, we split the latent code to two $256 \times 256 \times 16$ latent codes and feed them to two different decoders for appearance and geometry. The appearance decoder has two convolution layers with kernel size of 3. The geometry decoder has one convolution layer with kernel size of 3 and one convolution layer with kernel size of 1.

1.1.4. Deformer

We use a deformer to transform the avatar $\mathcal{G}$ from the canonical space into posed space. For each Gaussian primitive $\mathcal{G}_k$, the deformed Gaussian center and rotation matrix μ'_k and R'_k are computed as:

$$\mu'_k = T\mu_k, R'_k = TR_k, \text{ where } T = \sum_{i=1}^{n_b} w_i B_i.$$

Here n_b is the number of joints, B_i is the bone transformation matrix for joint $i \in \{1, ..., n_b\}$, and w_i is the skinning weight, which determines the influence of the motion of each joint on μ_k. We follow previous work AG3D [4] to represent the skinning weight as a low-resolution voxel grid. The skinning weights on the voxels are calculated by accumulating the weights of K nearest vertices on the SMPL-X surface. The contribution of vertices is inverse to the distance. The skinning weights of Gaussian primitives are obtained by trilinear interpolation from the voxel grid.

1.1.5. Training Details

Here we detail the training of the diffusion model. Given a latent code X_i, the latent diffusion model (LDM) adds Gaussian noise $\epsilon \in \mathcal{N}(0, I)$ into the latent code. At time step t, with noise schedule functions $\alpha(t)$ and $\sigma(t)$, the noisy code is $X_i^{(t)} = \alpha(t)X_i + \sigma(t)\epsilon$. To train the diffusion model, we use:

$$\mathcal{L}_{\text{diff}}(\{X_i\}, \phi) = \mathop{\mathbb{E}}_{i,t,\epsilon} \left[\frac{1}{2} w^{(t)} \left\| \hat{X}_i - X_i \right\|^2 \right]$$

*Equal contribution

where $\hat{X}_i = \hat{X}_\phi(X^{(t)}, t)$ is the denoised latent code with time step $t \sim \mathcal{U}(0, T)$, $\hat{X}_\phi$ represents the time-conditioned denoising network, $w^{(t)}$ is an empirical time-dependent weighting function, and ϵ is the added noise.

Following [2], the time-dependent weighting function $\omega(t) = (\alpha(t)/\sigma(t))^{2\omega}$, ω is an empirically chosen hyperparameter. We train the diffusion model from scratch and the latent code is randomly initialized for each subject.

1.1.6. Pose Estimation from Multi-view Images

After we obtain 6 synthetic human images, we use Openpose [1] to detect 25 2D keypoints from each image. We then triangulate the 2D keypoints to obtain 3D keypoints. Finally, we fit SMPL to the 3D keypoints.

1.1.7. Rendering Objective

Our rendering objective during inference is:

$$\begin{aligned}\mathcal{L}'_{\text{rend}}(\{X_i\}, \psi) =& \lambda_{\text{l2}}\mathcal{L}_{\text{l2}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \\ &+ \lambda_{\text{nc}}\mathcal{L}_{\text{nc}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}}.\end{aligned}$$

Here the $\mathcal{L}_{\text{l2}}$, $\mathcal{L}_{\text{vgg}}$, and $\mathcal{L}_{\text{reg}}$ are mentioned in the main paper. We use the normal consistency loss $\mathcal{L}_{\text{nc}}$ and depth distortion loss $\mathcal{L}_{\text{d}}$ from [8] to improve the geometry of the avatar. Specifically, $\mathcal{L}_{\text{d}}$ is formulated as:

$$\mathcal{L}_{nc} = \sum_i \omega_i (1 - n_i^T N) \tag{2}$$

where i indexes over intersected splats along the ray, ω denotes the blending weight of the intersection point, n_i represents the normal of the splat, and N is the normal estimated by the gradient of the depth map.

$\mathcal{L}_{\text{nc}}$ is formulated as:

$$\mathcal{L}_d = \sum_{i,j} \omega_i \omega_j |z_i - z_j| \tag{3}$$

where $\omega_i = \alpha_i \hat{\mathcal{G}}_i(\mathbf{u}(\mathbf{x})) \prod_{j=1}^{i-1} \left(1 - \alpha_j \hat{\mathcal{G}}_j(\mathbf{u}(\mathbf{x}))\right)$ is the blending weight of the i-th intersection and z_i is the depth of the intersection points. We use this loss to concentrate the weight distribution along the rays by minimizing the distance between the ray-splat intersections.

1.1.8. Image-Guided Sampling

Follow [2], we use the image-guided sampling method to provide a good initialization for the model fitting. Specifically, for a noisy code $X^{(t)}$ at time step t, we compute the approximated rendering gradient g with:

$$g \leftarrow \nabla_{x^{(t)}} \lambda_{rend} \sum_j \frac{1}{2} \left(\frac{\alpha^{(t)}}{\sigma^{(t)}}\right)^{(2\omega)} \mathcal{L}'_{rend}\left(\{\hat{X}_\phi(X^{(t)}, t)\}\right)$$

here the $\hat{X}_\phi(X^{(t)}, t)$ represents the time-conditioned denoising network. $(\alpha(t)/\sigma(t))^{2\omega}$ is a weighting factor based on the signal-to-noise ratio(SNR). The gradient g serves as an image-guided correction to the denoising output $\hat{X}^{(t)}$:

$$\hat{X}^{(t)} \leftarrow \hat{X}^{(t)} - \lambda_{gd} \frac{{\sigma^{(t)}}^2}{\alpha^{(t)}} g$$

with guidance scale λ_{gd}.

1.1.9. Pose Optimization

The generated synthetic images from multi-view diffusion models are inconsistent, and this results in inaccurate camera and body pose estimation. For better reconstruction results, we optimize the camera and body poses. For camera poses, we optimize the camera rotation, the elevation and azimuth angle of the camera position, and the distance between the object and the camera, we use the analysis by synthesis approach since the rendering process is differentiable with respect to the camera parameters. Specifically, we render the reconstructed avatar at different viewpoints, calculate photometric loss, then back propagate the loss to the camera parameters. For SMPL-X parameters, we optimize the global orient, translation, body pose, betas, and hand poses.

1.2. Implementation Details

1.2.1. Training Data

For fair comparison with other methods, following [7], our generative avatar model is trained on the THuman2.0 subset, which has 500 scans with ground truth SMPL-X parameters. For each object, we render 54 views rgb and normal images as ground truth. All images are rendered with 1024×1024 resolution. The cameras are set at $elev \in [-0.4, 0, 0.4]$ and $azim_i = 0 + 2\pi \frac{i}{18}$.

1.2.2. Training Time

For training the Avatar Prior, we use two NVIDIA Quadro RTX 6000 GPUs (24GB) over two days. Once trained, the Prior is used as a pre-trained model for fitting.

1.2.3. Camera

Current state-of-the-art multi-view diffusion models generate images from orthographic views. However, an orthographic camera is not yet supported for 2DGS [8]. To mitigate this, we simulate an orthographic camera by setting the focal length of the perspective camera to a very large value and placing it at a considerable distance. This approach effectively minimizes the perspective distortion, which enables us to use the results generated by multi-view diffusion models.

1.2.4. Mesh Extraction

Following [8], to extract meshes from reconstructed 2D splats, we render depth maps of the training views using the depth value of the splats projected to the pixels and utilize truncated signed distance fusion (TSDF) to fuse the reconstruction depth maps, using Open3D [16].

2. Evaluation Details

2.1. Test Data

Following PSHuman [10], we select 60 scans from the remaining scans in THuman2.1 and 60 scans from CustomHumans for quantitative evaluation. All samples in the THuman 2.0 and THuman 2.1 dataset are Asians, and samples from CustomHumans [6] dataset are more diverse. We use THuman 2.1 as an in-distribution test set and CustomHumans as an out-of-distribution test set. We collect in-the-wild images from the internet for qualitative comparison, most of our in-the-wild images are collected from freepik.com. For all test data, we use a foundation model [9] to extract the segmentation mask. At test time, all the SMPL-X parameters used are estimated instead of ground truth. For model fitting, the normals are generated by the multi-view diffusion model. When we calculate the normal loss, we normalize the magnitude of the rendered normal and estimated normal to 1.

2.2. Baseline

We compare our method with current state-of-the-art methods for single-view human reconstruction, including Human3Diffusion [13], SIFU [15], SiTH [7], and concurrent work PSHuman [10]. Human3Diffusion uses 3D Gaussian Splatting, but suffers from the low resolution of the multi-view diffusion model they use. Therefore, we only show qualitative results of Human3Diffusion. SIFU, SiTH, and PSHuman are mesh based methods.

2.3. Evaluation Metric

For appearance evaluation, we evaluate peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual similarity (LPIPS). We render color images from four viewpoints: $azim = 0, 90, 180, 270$ degrees relative to the input view for appearance evaluation. All metrics are calculated with 1024×1024 resolution.

For geometry comparison, we compute Chamfer Distance (CD), Point to Surface (P2S) distance, and normal consistency (NC). To avoid scale and depth ambiguity, we use ICP to align the predicted meshes to the ground-truth meshes before evaluation. The unit of CD and P2S in our results is cm.

For both appearance and geometry, we follow the same testing setup as SITH [7], and the SMPL-X pose is estimated, which is closer to real-world applications.

3. Additional Results

3.1. Ablations

Effect of pose optimization. Fig. 1 shows the importance of avatar prior in pose optimization. In this example, the initial pose estimation is inaccurate, due to self-occlusion, which causes noisy 2D joint detections. This results in unnatural reconstruction in the back region. By incorporating our generative avatar prior, this artifact can be corrected by optimizing pose with a more effective photometric rendering loss, leading to a more realistic avatar reconstruction.

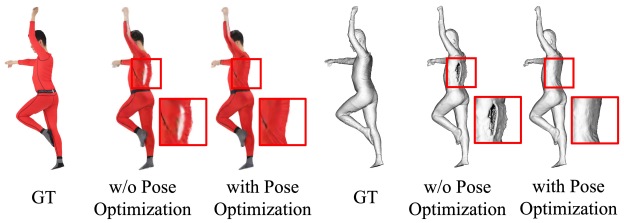

Figure 1. **Ablation of pose optimization.** The generative avatar prior helps correct inaccurate 3D pose estimation, leading to improved reconstruction quality.

Robustness to number of views. We present qualitative results using different numbers of input views. With a single input image, our generative avatar model can also reasonably reconstruct the 3D shape and appearance of novel identities. However, due to the limited availability of high-quality 3D training data, it may introduce artifacts on the backside and fail to capture details of unseen identities. To mitigate these limitations, in our method, we leverage our generative avatar prior as a complementary prior and integrate it with 2D diffusion models. This design enhances both appearance quality and 3D consistency.

Even with only front and back views, our method reconstructs plausible side views with minimal degradation in quality compared to reconstructions from six views. Here, we compare our two-shot approach with the state-of-the-art (SotA) method SiTH [7], which is designed for two-shot human reconstruction, as shown in Fig. 2. Our method achieves superior reconstruction quality, particularly in challenging regions such as hands, faces, and arms, with an improvement in side-view fidelity.

Moreover, as demonstrated in Fig. 2, our model benefits from an increasing number of input views, as it effectively aggregates multi-view information into a canonical space. Motivated by these findings, we leverage multi-view diffusion to hallucinate synthetic images, further enhancing reconstruction quality. Since our generative avatar model is robust to varying numbers of input views, it remains highly flexible and inherently compatible with other 2D diffusion models. This property also enables our method to integrate additional data sources, such as video sequences or video-based diffusion models.

3.2. Comparison with Gaussian Avatar

To further demonstrate the effect of our generative avatar prior on few-shot human reconstruction, we compare with

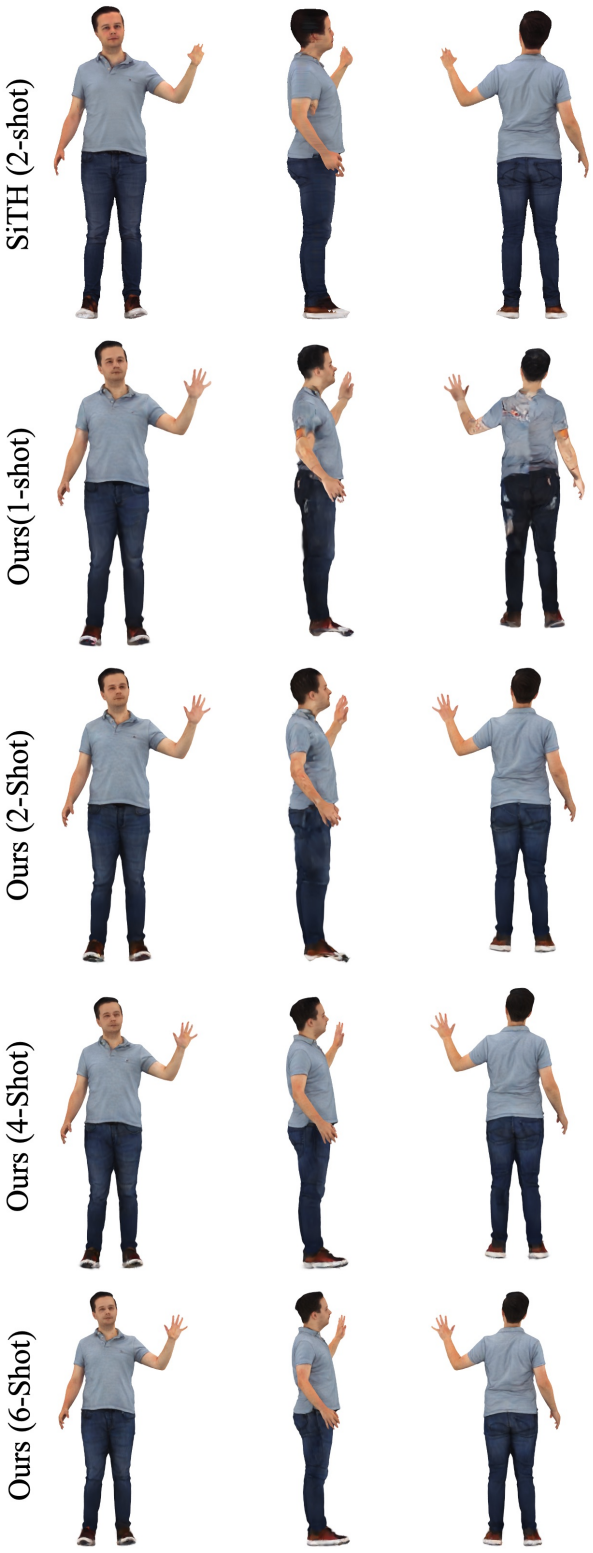

Figure 2. **Effect of number of views.** Our method is robust to the number of views used, even with only front and back views we can reconstruct reasonable side views.

SotA baselines, Animatable Gaussians [11], on the task of few-shot reconstruction. Here, for both of the methods, we leverage a pretrained multi-view diffusion model [10] to generate 6 synthetic images. Fig. 3 shows a qualitative comparison. Animatable Gaussians tend to reconstruct blurry appearance with abnormal colors, especially on the face region. This is because this method overfits to the sparse synthetic images, which are not inconsistent. However, with our generative avatar prior, our method can achieve large improvements in 3D consistency and appearance quality. As shown in Table 1, our method also outperforms [11] on all of the evaluation metrics.

Method	PSNR↑	SSIM↑	LPIPS↓
AnimatableGaussians [11]	21.234	0.921	0.092
Ours	**23.438**	**0.935**	**0.079**

Table 1. **Comparison with AnimatableGaussians [11] on CustomHuman Dataset.** Our method outperforms the SotA method on all of metrics for the task of few-shot reconstruction.

3.3. Unconditional Generation

Method	FID ↓	FID_{norm} ↓
PrimDiffusion [3]	68.60	NA
GETAvatar [14]	17.91	55.02
Ours	**15.59**	**25.63**

Table 2. We compare our method with other state-of-the-art methods on unconditional human generation.

To prove that our generative avatar prior is powerful enough, we conduct experiments to compare unconditional generation ability. Here, we choose two of the current state-of-the-art methods, GETAvatar [14] and PrimDiffusion [3] for quantitative and qualitative comparison. GETAvatar is a GAN based method, they use the tri-plane representation to model the avatar in canonical space, and then use an explicit mesh representation to model the avatar in deformed space. PrimDiffusion uses volumetric primitives to represent 3D Human body and they operate the denoising diffusion process on the volumetric primitives directly. All baselines are trained on Thuman 2.0 with 500 human scans. 50000 images are used to calculate the FID. Table 2 summarizes the generation result of our method compared to other methods. Overall, we perform better than other methods in all of the metrics. Fig. 4 shows that our generative method can generate more details and higher rendering quality, especially on face regions. Our geometry is also much better than other methods by producing more details like wrinkles. The im-

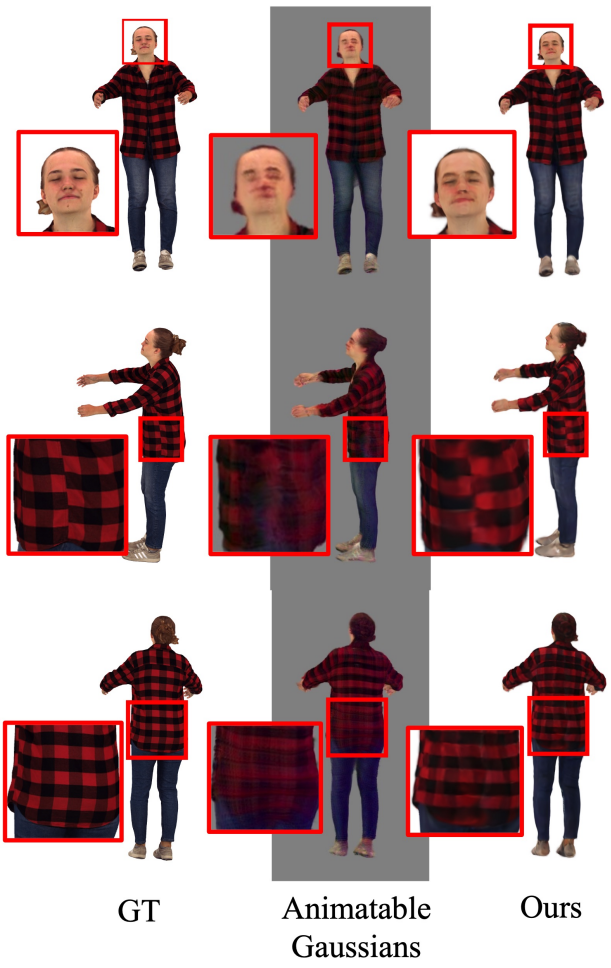

Figure 3. **Comparison with AnimatableGaussians [11].** Our method achieves better appearance quality.

provement stems from our exploration of a more powerful SMPL-anchored 2D Gaussian representation, which effectively captures both the geometry and appearance of avatars. SMPL body model provides initialization and enforces human body constraints, enhancing structural accuracy. Additionally, we introduce normal supervision during training, which significantly improves geometric fidelity.

3.4. Interpolation

Our generative avatar model learns the data distribution and enables interpolation between the training samples instead of just performing data retrieval. Here we show an interpolation result of our model in Fig. 5

3.5. Inconsistency of synthetic images.

Multi-view diffusion model generates sparse and inconsistent images, we provide some example results of [10] in Fig. 6 to show inconsistent face and human limb generation.

Figure 4. **Generation results.** Our model generates much more realistic appearance and geometry, especially on face regions.

Figure 5. **Interpolation results**.

Figure 6. **Inconsistency of synthetic images.**

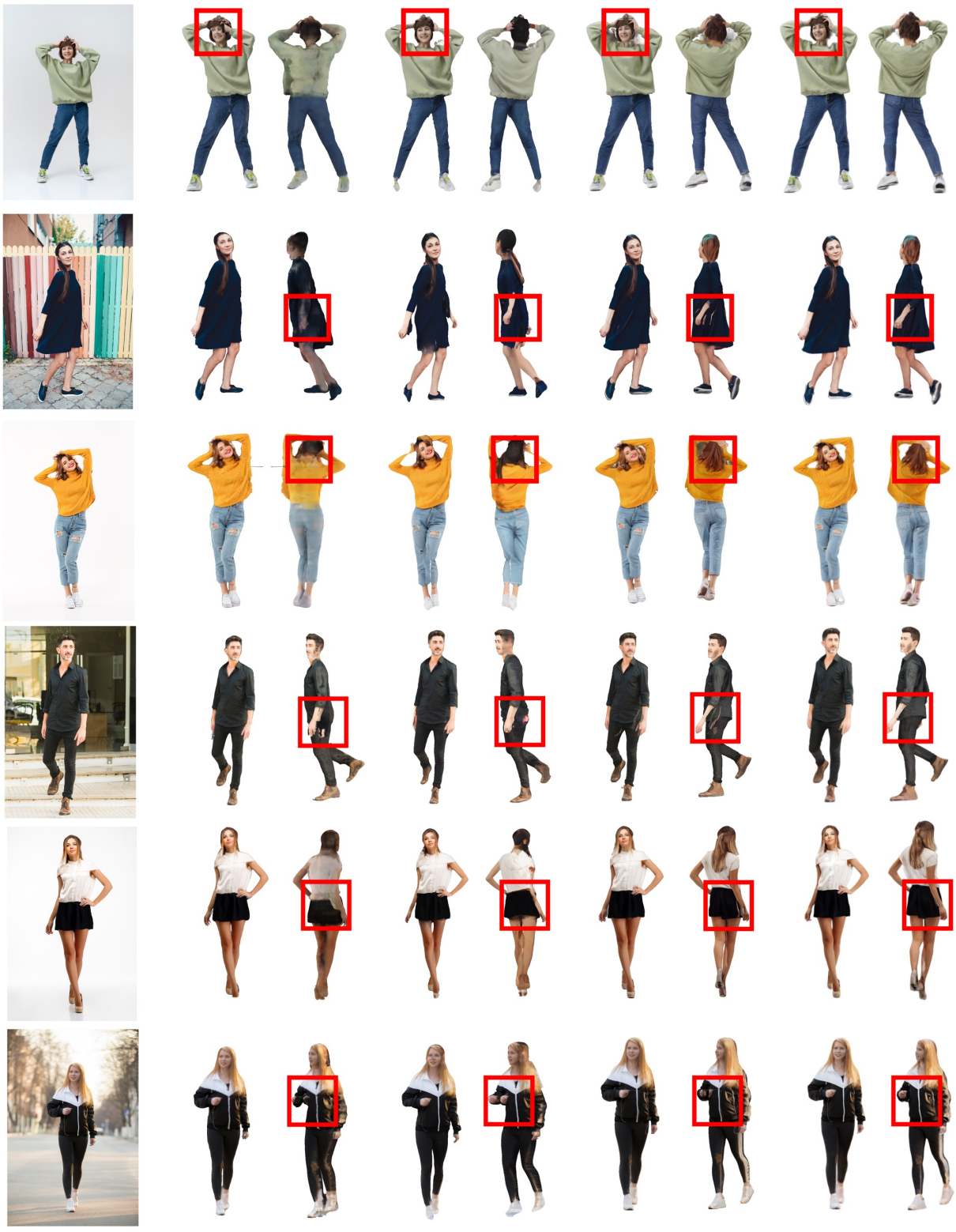

Figure 7. Qualitative comparison to SotA methods on in-the-wild-images.

3.6. Comparison to Baselines

In Fig. 7, we show additional comparisons between baselines and our method. Our method generates sharper images with more details and less artifacts. We handle self-occlusions well thanks to the 3D appearance prior from our model, while all other baselines suffer, especially in the sixth example. Concurrent method PSHuman [10] works better compared to other baselines, but struggles to model the areas between the arms and the head in the first and third examples. In contract, our method faithfully reconstruct all the challenging examples.

3.7. Additional Qualitative Results

We also show additional qualitative results to demonstrate our method generalizes well to all the in-the-wild scenarios. More results are available in the supplementary video.

4. Limitation and Future Work

Although achieving significant improvements, our method still has several limitations which are common to all existing methods.

Hand Pose: Our method tends to produce artifacts when the initial pose estimation is significantly incorrect. This issue is particularly pronounced for hands, which appear small in the image, making accurate hand pose estimation challenging. A potential solution is to incorporate a specially-designed hand pose estimator trained on specialized hand datasets, which could improve robustness and reduce artifacts in hand reconstruction.

Speed: Although already efficient, it takes almost 10 minutes to reconstruct a Gaussian avatar. This can further be solved by leveraging a more powerful image-conditioned diffusion [12] or image encoder [7] to initialize the fitting process.

Ambiguous Body Part Association: Our reconstructed avatar can be animatable without post-processing. However, under extreme unseen poses, it tends to generate artifacts, such as clothing patterns under the arms. This issue arises because the subject is observed in only a single pose, making it challenging to uniquely associate image observations with specific body parts, particularly in cases of occlusion or inter-part contact. A potential solution is to integrate a video diffusion model to synthesize multiple poses of the same subject (e.g., from a video), thereby improving robustness in handling occluded regions.

Loose Clothing: Similar to previous methods [10, 15], our method struggles to model realistic deformation of loose clothing such as skirts. These non-skeletal induced dynamics are beyond the scope of this work. Combining our model with physics-based simulation [5] can be a promising direction to explore.

Video Inputs: Our avatar model is defined in canonical space and can be reposed to fit a video input. To refine LBS weights, we can replace our deformer with AG3D [4] and optimize the weights during avatar reconstruction.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[2] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2416–2425, 2023. 2

[3] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023. 4

[4] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14916–14927, 2023. 1, 7

[5] Artur Grigorev, Michael J Black, and Otmar Hilliges. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2023. 7

[6] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. 3

[7] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 2, 3, 7

[8] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[9] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 3

[10] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human

Figure 8. Qualitative results on in-the-wild images

reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024. 3, 4, 5, 7

[11] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Figure 9. Qualitative results on in-the-wild images

Recognition, pages 19711–19722, 2024. 4, 5

[12] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 7

[13] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. 2024. 3

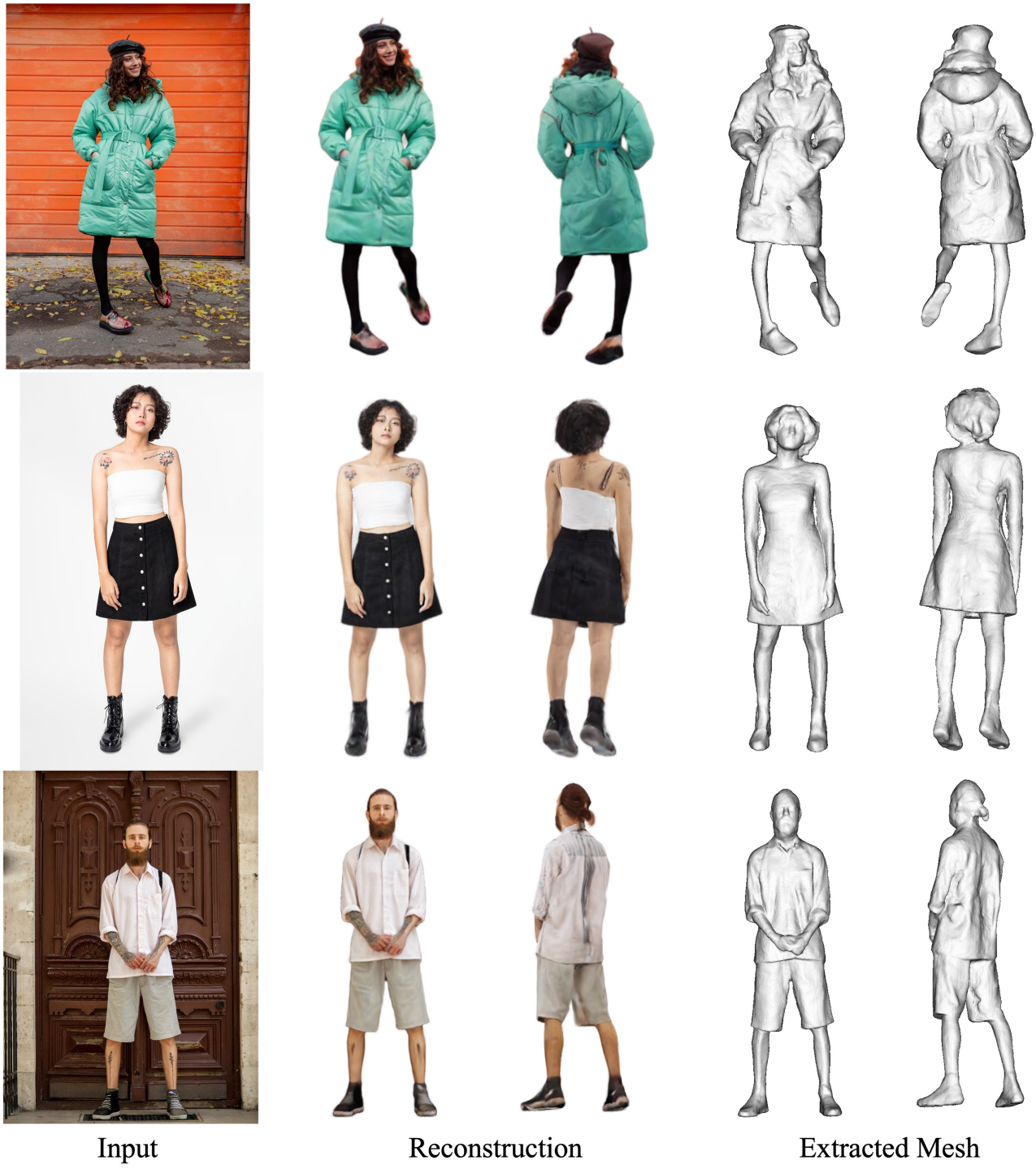

Figure 10. Qualitative results on in-the-wild images

[14] Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2273–2282, 2023. 4

[15] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the*

Figure 11. Qualitative results on in-the-wild images

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9936–9947, 2024. 3, 7

[16] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 2

Figure 12. Qualitative results on in-the-wild images

Figure 13. Qualitative results on in-the-wild images

Figure 14. Qualitative results on in-the-wild images

Figure 15. Animation results on in-the-wild images