Supplementary Material for *LayGA: Layered Gaussian Avatars for Animatable Clothing Transfer*

SIYOU LIN, ZHE LI, and ZHAOQI SU, Tsinghua University, China
ZERONG ZHENG, NNKosmos Technology, China
HONGWEN ZHANG, Beijing Normal University, China
YEBIN LIU, Tsinghua University, China

A1 IMPLEMENTATION DETAILS

A1.1 Model Architecture

Our model architecture is mostly the same as Animatable Gaussians [Li et al. 2024]. Please refer to their original paper for more details. We make a few modifications on top of their model: (1) For the single-layer model, we additionally add 2 channels to the outputs representing probabilities of p^{cloth} and p^{cloth} whether each Gaussian belongs to body or clothing. (2) We observe that the wavelet transform in StyleUNet leads to grid-like effect in the outputs. Since the geometry is also predicted by this network, there is a bias for the output point cloud to be scattered around the actual surface. For the single-layer model and the geometry layer in the multi-layer model, we replace all wavelet transforms with upsampling and downsampling layers to remove this bias. For the rendering layer prediction, we use a separate network that uses wavelet transform. (3) The input pose map is additionally channel-wise concatenated with a canonical spatial feature map $F \in \mathbb{R}^{H \times W \times 64}$ to increase the model's ability to represent spatially high-frequency changes. Here, $H = W = 256$ and the feature map is upsampled to 512×512 to match the size of input position maps. We set its value to 0 at initialization and do not optimize it until the 20000-th iteration. This feature map is considered part of model parameters and learned in an auto-decoding fashion, with a L_2 regularization term with weight 0.001. (4) We apply a small scale 0.2 to the network offset output for the single-layer model and the geometric layers of the multi-layer model. Moreover, the network predicts opacities, scales, rotations are predicted as residuals w.r.t. values initialized using the method in 3DGS [Kerbl et al. 2023].

A1.2 Dataset

We train our models on one sequence from the AvatarRex [Zheng et al. 2023] dataset and three sequences (the first sequence from subjects 02, 05 and 08) from the ActorsHQ [Işık et al. 2023] dataset. We also capture a new sequence of a man wearing tight white T-shirt dancing. For the AvatarRex sequence, we use 13 view for training. For the ActorsHQ sequences, we only use the 39 views that covers the full body. Furthermore, for Actor08, its clothing is loose and motion stochasticity becomes apparent if all frames are included. Thus, we only use the first 1700 frames in this sequence. Our own captured sequence contains 15 views and 800 frames. Since we have trained on nearly all frames of each sequence, for the quantitative evaluation, we use the first 1000 frames of a front-facing camera.

A1.3 Training Setup

We train our single-layer model for 200k iterations and our multi-layer model for 550k iterations. We set loss weights for the single-layer model as: $\mathcal{L}_{\text{off}} = 0.01$, $\mathcal{L}_{\text{TV}} = 1$, $\lambda_{\text{stitch}} = 0.5$, $\lambda_{\text{edge}} = 1000$, $\lambda_{\text{normal}} = 0.25$, $\lambda_{\text{label}} = 0.003$, $\lambda^{\text{label}}_{\text{TV}} = 0.01$, $\lambda^{\text{label}}_{\text{stitch}} = 0.003$, $\lambda_{\text{L1}} = 1$, $\lambda_{\text{ssim}} = 0.2$, $\lambda_{\text{lpips}} = 0.01$. For the multi-layer model, we set: $\mathcal{L}_{\text{normal}} = 0.01$, $\lambda_{\text{coll}} = \lambda_{\text{layer}} = 100$, $\lambda_{\text{cd}} = 0.5$. Other training setup are the same as Li et al. [2024]. Note that for fairness, we do not use parametric templates and view-dependent appearances for Li et al. [2024]. Furthermore, collision handling is not used for the evaluation in Sec. 4.1.

Given the predicted Gaussian maps M^{f}_{g} and M^{b}_{g}, we first extract the channels representing offsets $\Delta\bar{x}_i$ of the geometric layer to obtain offset maps $M^{\text{f}}_{\text{off}}$ and $M^{\text{b}}_{\text{off}}$. Let M_{off} be their concatenation channelwise. The TV loss $\mathcal{L}_{\text{TV}}$ is implemented as the average of $|M_{\text{off}}(i, j, c) - M_{\text{off}}(i', j', c)|^2$. The average is taken over all neighboring pixel pairs (i, j) and (i', j'), and over all channels c. For $\mathcal{L}_{\text{stitch}}$, we first add $M^{\text{f}}_{\text{off}}$ and $M^{\text{b}}_{\text{off}}$ to the base SMPL-X position map $M^{\text{f}}_{\text{pos}}$ and $M^{\text{b}}_{\text{pos}}$ to obtain deformed position maps $M^{\text{f}}_{\text{def}}$ and $M^{\text{b}}_{\text{def}}$. Then $\mathcal{L}_{\text{stitch}}$ is defined as the average of $\|M^{\text{f}}_{\text{def}}(i, j) - M^{\text{b}}_{\text{def}}(i, j)\|^2$ over all boundary pixels (i, j) of the template mask.

A1.4 Garment Segmentation masks

We use the SCHP [Li et al. 2020] model trained on the ATR dataset to obtained segmentation masks for garments. Since these masks are generally inaccurate, we use the binary masks provided by datasets to refine garment segmentation masks, according to the following rules:

- If a pixel is not valid in the binary, then it is considered as *background*;
- Otherwise, if it is labeled as background by SCHP, then it is considered as *undetermined*;
- Otherwise, if it is label as non-upper-clothes by SCHP, then it is considered as *body*;
- Otherwise, it is considered as *clothing*.

A2 EXTENDED EVALUATIONS AND DISCUSSIONS

A2.1 The Separation of Geometric and Rendering Layers

While we have introduced geometric constraints to allow Gaussians to reconstruct the smooth geometry for collision handling during clothing transfer, we empirically found geometrically constrained

Authors' addresses: Siyou Lin, linsy21@mails.tsinghua.edu.cn; Zhe Li, liz19@mails.tsinghua.edu.cn; Zhaoqi Su, suzq13@tsinghua.org.cn, Tsinghua University, 30 Shuangqing Road, Haidian, Beijing, 100084, China; Zerong Zheng, NNKosmos Technology, Hangzhou, Zhejiang, China, zrzheng1995@foxmail.com; Hongwen Zhang, Beijing Normal University, 19 Xinjiekouwaidajie Street, Haidian, Beijing, 100875, China, zhanghongwen@bnu.edu.cn; Yebin Liu, Tsinghua University, 30 Shuangqing Road, Haidian, Beijing, 100084, China.

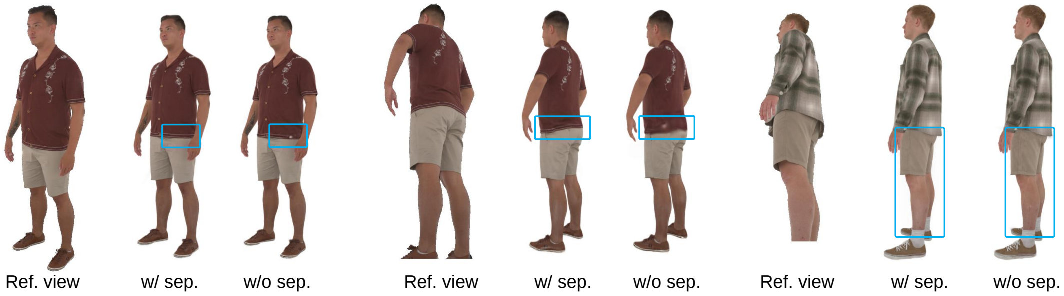

Fig. A1. Evaluation of the separation of geometric and rendering layers. Rendering results without using the separation scheme, i.e., directly rendering the geometric layer, show artifacts such as color flickering and blurred boundaries. Please zoom in for details.

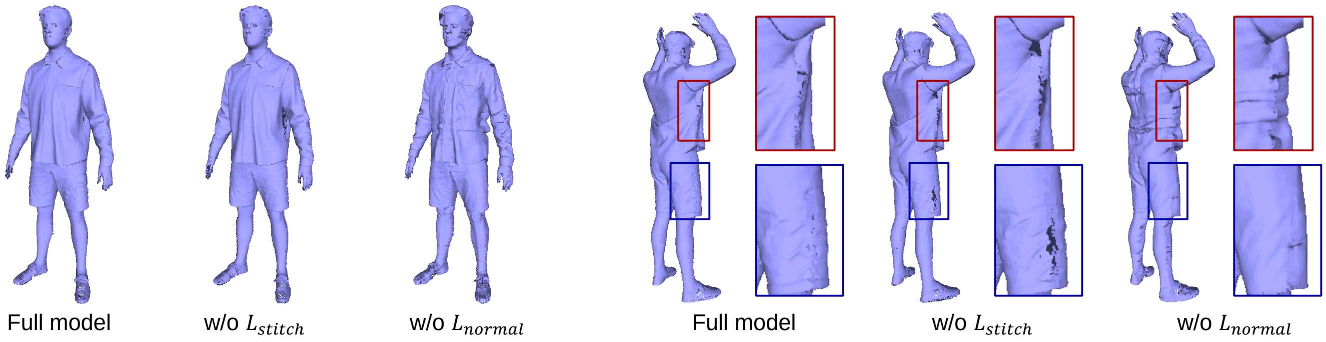

Fig. A2. Ablation studies on geometric constraints.

Gaussians may lower the rendering quality. Fig. A1 shows 3 groups of results rendered with and without the separation of geometric and rendering layers. Note that models without rendering layers may produce color flickering and blurred boundaries in a novel view. These artifacts are more obvious in our supplementary video. This justifies the use of additional rendering layers for improving the rendering quality. Note that the smooth geometric layers are still needed for modeling body-cloth relations for clothing transfer and collision handling.

A2.2 Ablation Study on the Geometric Constraints

In Fig. A2 we show the effects of $\mathcal{L}_{\text{stitch}}$ and $\mathcal{L}_{\text{normal}}$. Without $\mathcal{L}_{\text{stitch}}$, the reconstructed geometry may break open near the boundaries of front and map projection maps. Without $\mathcal{L}_{\text{normal}}$, the reconstructed geometry may become highly irregular due to shape-radiance ambiguity, i.e., a wrong geometry producing correct renderings for a training view. We remark that while $\mathcal{L}_{\text{normal}}$ and $\mathcal{L}_{\text{stitch}}$ are enough for a good reconstruction, we keep the other commonly used regularization terms $\mathcal{L}_{\text{off}}$, $\mathcal{L}_{\text{TV}}$ and $\mathcal{L}_{\text{edge}}$ as a safeguard to penalize overly large distortions, and to make the model favor small deformations.

A2.3 Ablation Study on the Collision Resolving Strategy

As emphasized before, the key to successfully transferring clothing to different body shapes is the collision handling strategy using the geometric layers of our model. To evaluate the necessity of these technical choices, we perform clothing transfer using a few alternatives. Let A and B be two subjects and we wish to transfer the clothing of A to B. Fig. A2 shows the results obtained by: (a) our full model; (b) directly using Eq. (16) as new garment Gaussian positions without subsequent collision handling (no collision handling); (c) computing ξ by directly adding the offsets predicted by $\mathcal{F}_A^{\text{cloth}}$ to the SMPL-X template of B instead of Eq. (16); (d) directly using the rendering layers in Eq. (16) to derive new positions; (e) directly using the outputs from two models without any post-processing.

Results in Fig. A3 show that only our full model is capable of properly handling collisions. (b) indicates that the Laplacian-based update steps are necessary for the clothing to remain a valid shape. (c) suggests that using relative offsets, i.e., Eq. (16) is important. (d) exhibits the necessity of utilizing the geometric layers for handling collision. These results thus validate our technical choices in modeling layered avatars.

A2.4 More Clothing Types

Pants. While we have primarily focused on using upper clothes as the outer layer and other parts as the inner layer, our method can be used to transfer any garment that is the outmost layer, including pants. Fig. A4 (left) shows the result of transferring pants to a different subject. Note that to avoid collision between the pants and the shirt, we train 3 models in this specific case: (1) a layered model for the subject providing pants, with pants as the outer layer; (2) a layered model for the subject providing shirt and body, with the

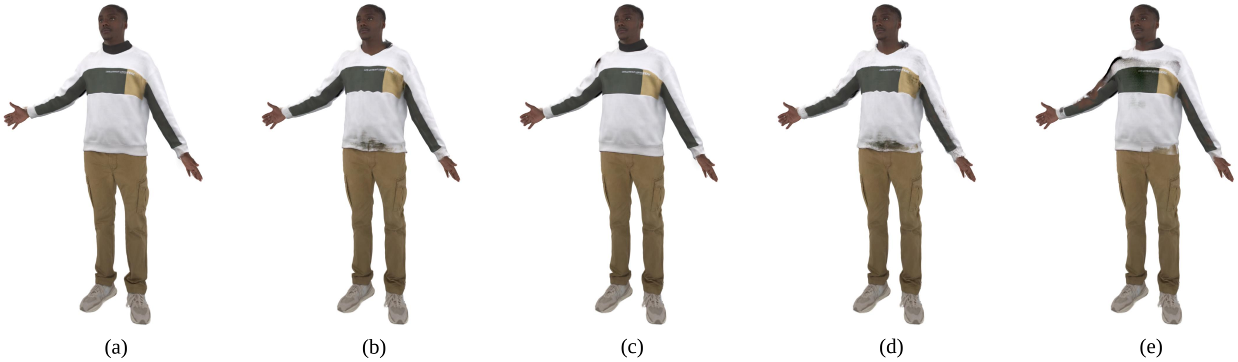

Fig. A3. Ablation studies. (a) Our full model; (b) Directly using Eq. (16) as new garment Gaussian positions without subsequent collision handling (no collision handling); (c) Computing ξ by directly adding the offsets predicted by $\mathcal{F}_A^{\text{cloth}}$ to the SMPL-X template of B instead of Eq. (16); (d) directly using the rendering layers in Eq. (16) to derive new positions; (e) Directly using the outputs from two models without any post-processing.

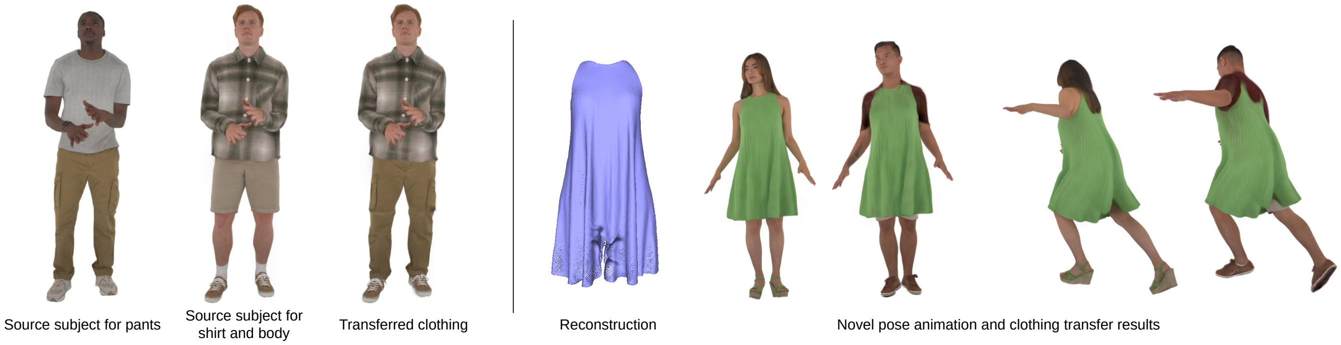
Source subject for pants
Source subject for shirt and body
Transferred clothing
Reconstruction
Novel pose animation and clothing transfer results

Fig. A4. Results of more clothing types.

shirt as the outer layer; (3) another layered model for the second subject, but with both the shirt and his pants as the outer layer. During transfer, we first dress the pants of model 1 to the body of model 3, and then dress the shirt of model 2 to it.

Skirts. Very loose clothing such as skirts remain a limitation. The main obstacle for modeling skirts is the topological discrepancy between the base SMPL-X model and skirts. When applied to skirt data, the reconstructed geometry is not continuous between the legs as shown in Fig. A4 (right). Consequently, the lower part of skirt becomes blurry or even break open when animated.

A2.5 Runtime statistics

For generating a single frame on an RTX4090, the model forward step costs 0.2s, collision handling costs 10 20s, 3DGS rendering costs 0.01s (RTX4090). Note that collision handling is solved with scipy on CPU because PyTorch doesn't provide an interface for GPU-based sparse least squares. A GPU-based solver may significant accelerate this step but this is left as future work.

REFERENCES

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *TOG* 42, 4 (2023), 1–12.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *TOG* 42, 4 (2023), 1–14.

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2020), 3260–3271.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2024. 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 Recognition (CVPR)*.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023. AvatarReX: Real-time Expressive Full-body Avatars. *TOG* 42, 4 (2023).