

GeoAvatar: Geometrically-Consistent Multi-Person Avatar Reconstruction from Sparse Multi-View Videos

Supplementary Material

1. Preprocess Details

1.1. Segmentation Map Estimation

Segmentation maps are estimated using SAM2 [5], a precise segmentation model designed for video inputs. A few pixel points in the first frame are manually selected to indicate objects of interest, after which SAM2 generates a segmentation map for the frame and tracks the segmented objects across all subsequent frames.

SAM2 often struggles with fine details such as fingers and contact regions. To mitigate these issues, we apply a guided filtering technique to refine the segmentation mask. This improves boundary accuracy by fitting segmentation maps to the boundaries observed in the frame images.

1.2. Keypoint Estimation

Keypoints of each object are estimated using Sapiens [2], a human-centric vision model trained on large-scale datasets. Since Sapiens is optimized for single-object images, its multi-object keypoint estimation is unreliable. To address this, each object is first cropped in every frame based on the segmentation maps obtained from SAM2. Then, Sapiens estimates the keypoints for each cropped object accurately even in heavily occluded regions.

1.3. Pose Estimation

Recent off-the-shelf pose estimation models [1, 6] support multi-object tracking from a monocular video with predicting camera parameters. However, these models are not well-suited for sparse view datasets where calibrated camera parameters are provided. Additionally, in close interaction scenes with heavy occlusion, tracking becomes less reliable. To address these challenges, we select a multi-view consistent pose estimation model MultiviewSMPLify-X [4, 7] as our baseline. This model is designed to estimate SMPL-X parameters of a single-object from multi-view images, and we extend it to multi-object multi-view videos.

Learning SMPL-X parameters for multiple objects simultaneously destabilizes the training process. The first frame in a video sequence typically consists of A-posed objects without close interactions, making pose estimation easier. Starting with the first frame, we perform pose estimation, and for subsequent frames, we use the predicted parameters from the previous frame as initialization. To ensure temporal consistency, we introduce a loss that measures the difference between the optimized SMPL-X mesh vertices from the previous frame and the predicted SMPL-X mesh

vertices in the current frame. This temporal consistency mechanism helps stabilize training, particularly in views with heavy occlusion due to close interactions, where the information from prior frames proves beneficial for maintaining reliable learning.

1.4. Pose Refinement

The accuracy of the pose estimation model is crucial for the quality of the reconstructed avatars. Thus, we refine the pose estimation results using surface ordering loss to reduce penetrations occurring in SMPL-X meshes. We render the segmentation map of SMPL-X meshes and compute the surface ordering loss. The regularization term of SMPL-X parameter is also added to prevent the SMPL-X mesh from being too deviated. This refinement enhances the quality of the reconstructed avatars by reducing the penetrations between the SMPL-X meshes.

2. Details of Auxiliary Deformation Fields

GART [3] proposes to model deformation of clothes, which is the limitation of SMPL-X-based deformation, by adopting latent bones and achieves fast training within 5 minutes. Thus, we follow GART to deform our Gaussian avatar with given per-frame SMPL-X parameters. GART represents deformation using the SMPL human avatar template with linear blend skinning (LBS). While SMPL-X effectively represents the geometry and deformation of the human body, it has limitations in capturing the geometry and deformation of clothed humans. To address this, GART introduces optimizable latent bones and their corresponding learnable LBS weight, which are represented by MLP, to model the deformation of clothing that is distant from the human body. Additionally, GART learns a 3D grid volume that corrects LBS weights, allowing for the representation of non-rigid motion.

References

- [1] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, pages 14783–14794, 2023. 1
- [2] R. Khrodar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024. 1
- [3] J. Lei, Y. Wang, G. Pavlakos, L. Liu, and K. Daniilidis. Gart: Gaussian articulated template models. In *CVPR*, pages 19876–19887, 2024. 1

- [4] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands and face and body from a single image. In *CVPR*, 2019. [1](#)
- [5] N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [1](#)
- [6] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J Black and. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *CVPR*, pages 8856–8866, 2023. [1](#)
- [7] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, pages 1–1, 2021. [1](#)