

Training-free Multi-view 4D Human Motion Reconstruction Virtual Reality System — Supplementary Material

Yijie Li¹, Ce Zheng^{1,†}, Yijie He¹, Joel Julin¹, Ryosuke Ichikari², Satoki Ogiso², Satoshi Nakae², Akihiro Sato², Takeshi Kurata², László A. Jeni^{1,†}

¹ Carnegie Mellon University

² National Institute of Advanced Industrial Science and Technology

[†]Corresponding authors

A. Overview

The supplementary material is organized into the following sections:

- Section B: Distinction of our multi-view multi-person human mesh recovery approach
- Section C: Ablation study.
- Section D: More visualization results.
- Section E: Additional quantitative results.
- Section F: Theoretical processing time analysis

B. Distinction of our multi-view multi-person human mesh recovery approach

For the human recovery component, we emphasize the strong generalization ability of our approach. As shown in our demo video (start at 3:50), our system achieves accurate multi-view 3D human mesh reconstructions on self-collected, in-the-wild sequences **without any training on these videos**. In contrast, existing single-view methods (e.g., Multi-HMR[1], SA-HMR [4]) perform poorly in such settings due to inherent depth ambiguities. Current multi-view methods are either limited to skeleton-only outputs (e.g., MVP[6], VoxelPose[5], TEMPO[2]) or constrained to handling only up to two people (e.g., AvatarPose [3]). Our framework supports full human mesh recovery for more than two individuals, as evidenced by the example in our demo (start at 0:14) involving 3–4 people. **Most importantly, existing single-view and multi-view methods typically cannot be directly applied to new scenarios, as they require dataset-specific fine-tuning to perform reliably.** In contrast, our system is plug-and-play and generalizes well to unconstrained scenes. Our focus is on enabling practical, generalizable deployment in real-world applications with less adaptation effort.

C. Ablation Study

In this section, we provide multiple ablation study results on pipeline components, backbone network type, and the number of camera views on the CHI3D and Hi4D datasets.

C.1. Number of camera views

To further analyze the impact of the number of camera views on our multi-view pipeline, we evaluate our approach using 2, 3, 4, and 5 views on the Hi4D dataset, as shown in Table 1. The results show that three mesh-related metrics: MPJPE, PA-MPJPE, and PVE improve as the number of camera views increases. As for the translation error, it significantly decreased from 72.08 mm to 59.07 mm when we increased the setting from 2 camera views to 5 camera views.

Table 1. Ablation study on different numbers of camera views on the Hi4D dataset

Ablation	Hi4D			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓
Ours (2 camera)	87.23	56.43	101.61	72.08
Ours (3 camera)	83.30	53.34	96.20	66.67
Ours (4 camera)	83.03	53.14	95.10	61.74
Ours (5 camera)	83.20	52.81	94.47	59.07

C.2. Model components

C.2.1. Multi-view mesh refinement

To refine the mesh prediction in a training-free and optimization-free manner, we propose to use re-projected MPJPE as a metric to select the best SMPL/SMPL-X prediction directly from single-view prediction. Let $\hat{\mathbf{X}}^{(i)} \in \mathbb{R}^{J \times 3}$ be the 3D joints predicted in camera i coordinates, with extrinsics $(\mathbf{R}_i, \mathbf{t}_i)$. The world-frame candidate is

$$\hat{\mathbf{X}}_w^{(i)} = \mathbf{R}_i^\top (\hat{\mathbf{X}}^{(i)} - \mathbf{t}_i^\top). \quad (1)$$

Table 2. Ablation study on multi-view SMPL/SMPL-X mesh refinement. MV-MR is the multi-view mesh refinement module.

Ablation	Hi4D				CHI3D			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓
Ours (w/o MV-MR)	92.39	59.36	108.46	60.68	64.33	42.93	69.09	39.75
Ours	83.20	52.81	94.47	59.07	56.13	38.22	58.36	39.60

Table 3. Ablation study on re-centering the joints around the pelvis.

Ablation	Hi4D				CHI3D			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓
Ours (w/o Re-centering)	83.20	52.81	94.47	59.07	68.63	42.39	71.53	87.93
Ours (w/ Re-centering)	80.18	52.52	91.86	67.04	56.13	38.22	58.36	39.60

Reprojecting into another camera v gives

$$\hat{\mathbf{X}}_c^{(i \rightarrow v)} = \mathbf{R}_v \hat{\mathbf{X}}_w^{(i)} + \mathbf{t}_v^\top. \quad (2)$$

The reprojected multi-view MPJPE for candidate i is

$$\mathcal{E}(i) = \frac{1}{MJ} \sum_{v=1}^M \sum_{j=1}^J \|\hat{\mathbf{x}}_{c,j}^{(i \rightarrow v)} - \hat{\mathbf{x}}_j^{(v)}\|_2. \quad (3)$$

Finally, we select the refinement with the lowest error:

$$i^* = \arg \min_i \mathcal{E}(i), \quad \hat{\mathbf{X}}_w^* = \hat{\mathbf{X}}_w^{(i^*)}. \quad (4)$$

This method can not only aggregate the best human mesh from multi-view prediction, but also can naturally alleviate the occlusion issues by selecting the mesh from another view without severe occlusion. The basic idea of this module is: the prediction from the view with occlusion usually have bad re-projected MPJPE on other views, which is unlikely to be selected through Equation C.2.1.

C.2.2. Results

Considering that each of our model components is irremovable, it is difficult to conduct an ablation study on model components. However, we can analyze the effectiveness of the multi-view mesh refinement stage of our approach by comparing it with the version without mesh refinement, as shown in Table 2. The results show that on the Hi4D dataset, applying our multi-view mesh refinement can significantly decrease the MPJPE from 92.39 mm to 83.20 mm and the PVE from 108.46 mm to 94.47 mm. On the CHI3D dataset, a substantial improvement is also observed. Besides, we compare the results of performing re-centering and without re-centering before flattening the joints, as shown in Table 3. In our approach, we take the flattened 3d joints as features for K-Means multi-view matching. The estimation of translation from the single-view HMR model

may affect the matching quality. The results show that on both the Hi4D and CHI3D datasets, applying re-centering on body joints can significantly improve the mesh recovery performance. However, for the translation error metric, using re-centering can result in a decrease in translation on the Hi4D dataset, but it can significantly improve the prediction on the CHI3D dataset.

C.3. Model backbone size

We also conduct additional experiments on both the Hi4D and CHI3D datasets using different backbones, ViT-small and ViT-large, as shown in Table 4. On both datasets, increasing model size from ViT-small to large has a significant impact on all 4 metrics.

D. More Visualization

D.1. Visualization for environmental reconstruction

As demonstrated in Figure 2, our system effectively utilizes 3D Gaussian Splatting for high-quality environmental reconstruction. The results highlight the impressive visual quality and fidelity of the reconstructed scenes, enabling seamless integration of the reconstructed scenes with human mesh recovery for immersive XR applications.

D.2. Visualization for integration of human mesh and 3D Gaussian Splatting scene.

As shown in Figure 1, our approach enables seamless integration of recovered human motion and 3D Gaussian Splatting scene.

D.3. Multi-view mesh recovery on Hi4D and CHI3D datasets

We present the qualitative results of our method on the CHI3D and Hi4D datasets, as illustrated in Figures 3 and 4. The samples in the figures include occlusion in some cam-

Table 4. Ablation study on different backbone size

Ablation	Hi4D				CHI3D			
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓
Ours (ViT-small)	83.2	52.8	94.5	59.1	56.1	38.2	58.4	39.6
Ours (ViT-large)	61.9	46.4	78.6	38.7	49.7	31.0	51.2	34.3



Figure 1. Visualization of human mesh and motion recovered from Shelf sequence and inserted into Garden 3D Gaussian-Splatting scene.

Table 5. Quantitative comparison with AvatorPose on the CHI3D dataset. The backbone network of our approach is fully fine-tuned on the training set.

Methods	CHI3D				
	MPJPE ↓	PA-MPJPE ↓	PVE ↓	Transl. ↓	Recall ↑
AvatorPose [3]	32.98	-	-	-	99.85
Ours	28.13	20.12	31.35	18.10	99.86

era views, which may result in incorrect predictions of overlapping humans if using a single-view approach. However, in our approach, the multi-view and MPJPE metric-based mesh refinement successfully alleviates this issue.

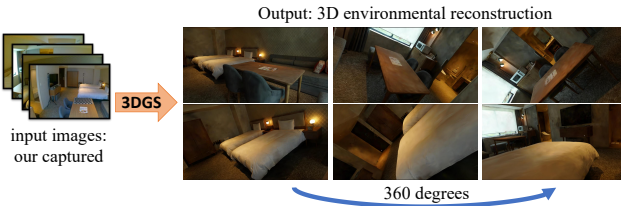


Figure 2. Visualization of environmental reconstruction given input images. Please refer to our demo video for the visual results.

E. Additional Quantitative Results

E.1. Fine-tuning single-view Multi-HMR

To clarify the results on the CHI3D dataset in Table 2 of our main paper and Table 5, we will provide the fine-tuning details in this section. In our main paper, we call the results of our approach on the CHI3D dataset with fine-tuning as partially fine-tuned results, which only take the data from one camera view in training. However, to enable fair com-

parison with AvatorPose [3], we perform a full-finetuning that uses data from all 4 camera views. We fine-tune the Multi-HMR model for 10 epochs with a batch-size of 8 and using the Adam optimizer with a constant learning rate of $1e-05$.

E.2. Results

In Table 5, we compare our approach with AvatorPose [3], an optimization-based multi-view HMR approach. The results show that our approach has lower MPJPE and better recall, which demonstrates the effectiveness of our approach.

F. Theoretical processing time analysis

Inheriting the efficiency of Multi-HMR, our system can estimate multiple human meshes simultaneously with minimal impact on processing time. Specifically, for a single person, the small version of Multi-HMR-S requires approximately 31 ms to output a human mesh. Even when scaling up to multiple people, the processing time remains almost unchanged, demonstrating the scalability and efficiency of the approach. In our experiments, the number of people typically remains below 10, ensuring that the system operates efficiently. As shown in Table 6, we report the theoretical processing time per frame, comparing various components and view configurations. The overall processing time remains efficient, even with the addition of multi-view triangulation, human tracking, and motion smoothness optimization. While generating meshes using Multi-HMR-S is the most time-consuming step, the additional operations, such as person identification, mesh optimization, and tracking, are relatively fast, each requiring

	Mesh output (Multi-HMR)	Person identification across views	Mesh optimization	Human tracking	Motion smoothness	Overall
Fast-version (3-view)	31.61 ms	10.26 ms	13.24 ms	0.66 ms	0.82 ms	56.59 ms
Basic-version (3-view)	31.61 ms	10.26 ms	135.67 ms	0.66 ms	0.82 ms	179.02 ms
Fast-version (4-view)	31.86 ms	11.65 ms	17.26 ms	0.66 ms	0.82 ms	62.25 ms
Basic-version (4-view)	31.86 ms	11.65 ms	174.97 ms	0.66 ms	0.82 ms	219.96 ms

Table 6. Theoretical processing time per frame comparison across different components and view configurations.

only a few milliseconds. The “fast-version” refers to using fewer iterations (10) in the mesh optimization process to achieve higher processing speed, while “basic-version” indicates a greater number of iterations (100), producing more accurate meshes at the cost of increased processing time. All processing times were measured on a single NVIDIA Tesla V100 GPU. This highlights our system’s robustness and scalability, suitable for real-world applications requiring fast and accurate 3D mesh recovery and tracking.

References

- [1] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. [1](#)
- [2] Rohan Choudhury, Kris M. Kitani, and Laszlo A. Jeni. Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *ICCV*, 2023. [1](#)
- [3] Feichi Lu, Zijian Dong, Jie Song, and Otmar Hilliges. Avatar-pose: Avatar-guided 3d pose estimation of close human interaction from sparse multi-view videos. In *European Conference on Computer Vision*, pages 215–233. Springer, 2024. [1](#), [3](#)
- [4] Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Learning human mesh recovery in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17038–17047, 2023. [1](#)
- [5] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *ECCV*, 2020. [1](#)
- [6] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34:13153–13164, 2021. [1](#)

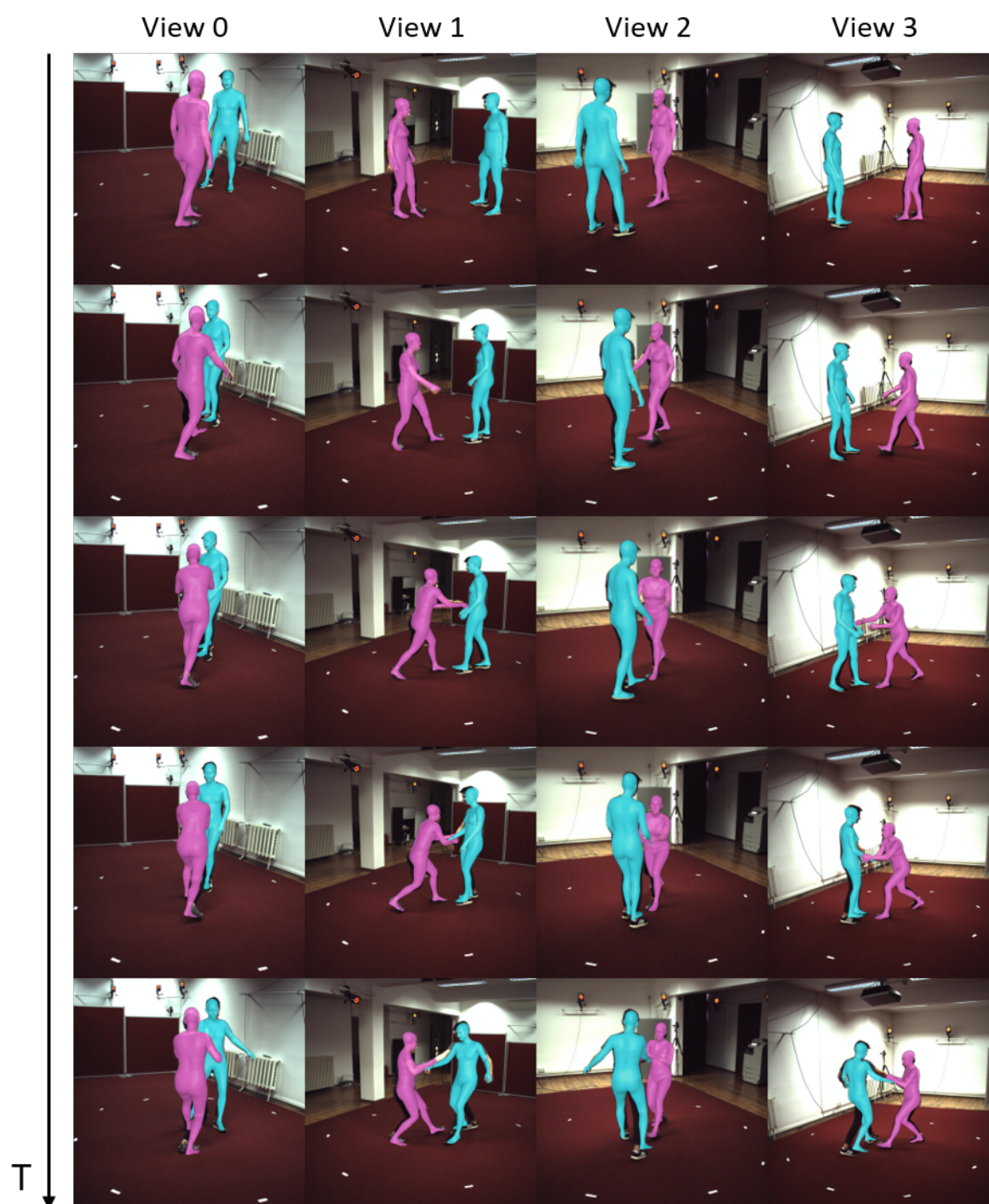


Figure 3. Visualization of human mesh and motion recovered from the CHI3D dataset.

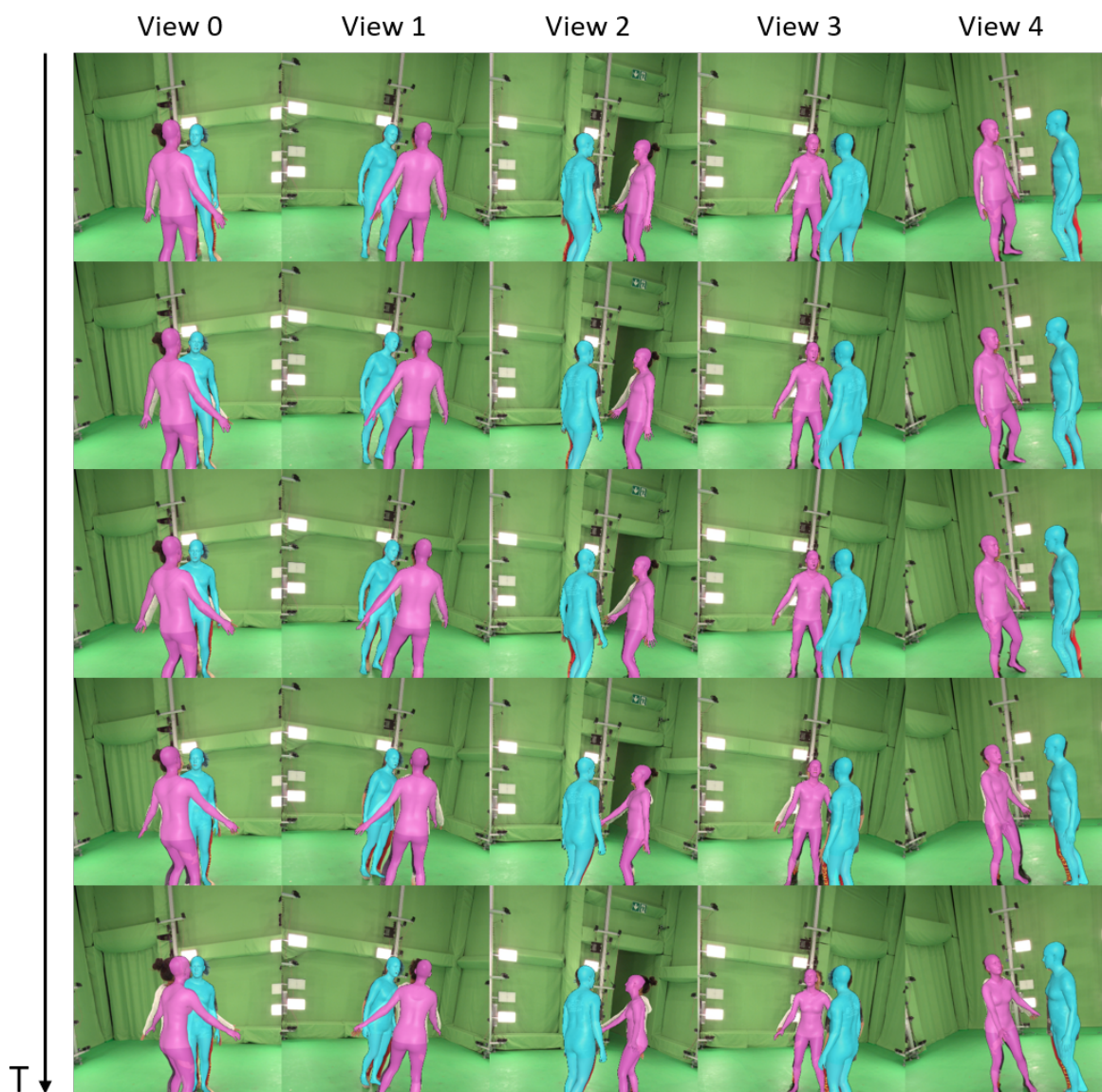


Figure 4. Visualization of human mesh and motion recovered from the Hi4D dataset.