

Link to the Past: Temporal Propagation for Fast 3D Human Reconstruction from Monocular Video

Supplementary Material

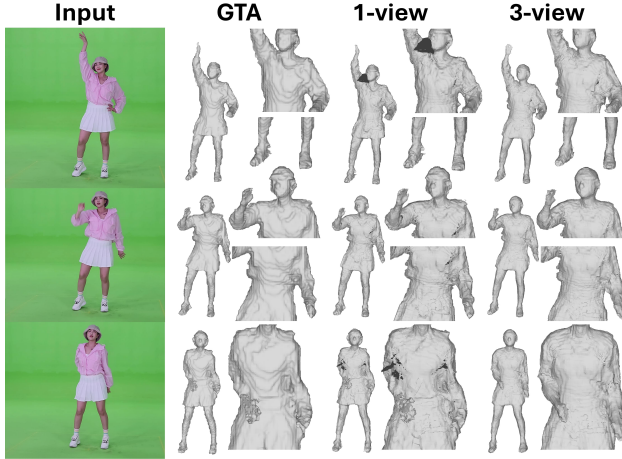


Figure 7. Qualitative comparison on clean-background video (for clear comparison) between baseline, single-frame reconstruction, and our three-view approach by merging canonical shapes.

6. Implementation Details

Training and Inference Setup. All comparative experiments in Table 2 were conducted on a single NVIDIA RTX 4090 GPU. Training times represent the duration required to train each method only on the “bike” sequence from NeuMan dataset [15]. Methods without implementation support for NeuMan dataset (e.g., 3DGS-Avatar [28]) were excluded from training time comparison. We report average FPS across the entire sequence rather than maximum FPS, as our method exhibits speed variation—requiring a slower warm-up period for the first 5 frames (full reconstruction) followed by faster processing for subsequent frames. Other video-based methods generally maintain consistent processing times throughout the sequence. In contrast, our ablation studies (Table 3) report maximum FPS to demonstrate the peak capability of our acceleration strategies after the initial warm-up period. Only our method’s reported FPS in Table 2 includes both shape reconstruction and color inference.

Hyperparameter n . The frame threshold parameter $n = 5$ represents a practical balance between reconstruction quality and computational efficiency. This value was selected based on our observations during development and is supported by our experimental results in Table 6 and Figure 11, which show diminishing quality returns beyond 6-7 views. While our multi-view experiments used evenly-spaced orthogonal viewpoints to evaluate the method’s theoretical capabilities, the principle of canonical shape convergence ap-

Method	K	THuman2.0		
		Chamfer ↓	P2S ↓	Normal ↓
TPF3D-SIFU	1	0.5253	0.4422	0.0386
TPF3D-SIFU	2	0.5188	0.4415	0.0380
TPF3D-SIFU	3	0.5089	<u>0.4420</u>	<u>0.0375</u>
TPF3D-SIFU	4	0.5063	0.4421	0.0374
TPF3D-SIFU	5	0.5048	0.4434	0.0374
TPF3D-SIFU	6	<u>0.5039</u>	0.4449	<u>0.0375</u>
TPF3D-SIFU	7	0.4995	0.4461	0.0376

Table 4. Comparing the impact of **K** number of neighbors in coordinate mapping (Section 3.2) for **single-frame reconstruction**.

Method	K	THuman2.0		
		Chamfer ↓	P2S ↓	Normal ↓
TPF3D-SIFU-3v	1	0.4407	0.3596	0.0328
TPF3D-SIFU-3v	2	0.4240	0.3576	0.0321
TPF3D-SIFU-3v	3	0.4184	<u>0.3581</u>	0.0315
TPF3D-SIFU-3v	4	<u>0.4162</u>	<u>0.3581</u>	0.0313
TPF3D-SIFU-3v	5	0.4144	0.3590	0.0313
TPF3D-SIFU-3v	6	0.4179	0.3601	<u>0.0314</u>
TPF3D-SIFU-3v	7	0.4182	0.3623	0.0315

Table 5. Comparing the impact of **K** number of neighbors in coordinate mapping (Section 3.2) for **three-frame reconstruction**.

plies similarly to sequential frames in video as shown in Figure 7. The value $n = 5$ provides sufficient initial frames to establish a robust canonical human shape while allowing the system to transition to the more efficient inference mode quickly enough to increase the inference speed. This parameter can be adjusted based on specific application requirements.

Impact of **K in Coordinate Mapping.** The number of neighbors (**K**) in our coordinate mapping affects the trade-off between transformation smoothness and local detail preservation. We compare the results from single-view reconstruction and three-view reconstruction which we report in Table 4 and Table 5, respectively. Empirical evaluation shows steady improvement from $K=1$ to $K=5$ for chamfer distance and normal consistency while the P2S score decreases. Larger **K** values ($K > 5$) show diminishing returns and eventual degradation in performance. While these dif-

Method	Num. Views	THuman2.0		
		Chamfer ↓	P2S ↓	Normal ↓
TPF3D-SIFU	1	0.5047	0.4432	0.0374
TPF3D-SIFU	2	0.4982	0.4444	0.0368
TPF3D-SIFU	3	0.4144	0.3590	0.0313
TPF3D-SIFU	4	0.4223	0.3632	0.0318
TPF3D-SIFU	5	0.4147	0.3545	0.0310
TPF3D-SIFU	6	0.4056	0.3480	0.0305
TPF3D-SIFU	7	0.4027	0.3436	0.0303
TPF3D-SIFU	9	0.4003	0.3439	0.0303
TPF3D-SIFU	10	0.4035	0.3412	<u>0.0302</u>
TPF3D-SIFU	12	0.4009	0.3417	<u>0.0302</u>
TPF3D-SIFU	18	0.3970	<u>0.3403</u>	<u>0.0302</u>
TPF3D-SIFU	36	<u>0.4003</u>	0.3396	0.0301

Table 6. **Impact of view count on reconstruction quality.** We compare the geometric accuracy improvements by combining results from multiple-views on the THuman2.0 dataset [40]

ferences are measurable quantitatively, the visual variations in the final reconstruction are subtle, primarily noticeable in the texture creases becoming more defined as K increases, as shown in Figure 8. We adopt $K=5$ as our default setting based on these results.

Number of views. We analyze the relationship between viewpoint multiplicity and reconstruction quality in Table 6. Using the THuman2.0 [40] dataset, we evaluate configurations ranging from single to 36-view reconstructions, with viewpoints distributed at maximal angular separations (e.g., 0° , 180° for two views; 0° , 120° , 240° for three views; 0° , 90° , 180° , 270° for four views). Our analysis reveals consistent improvements in geometric accuracy with additional viewpoints up to 7 views, beyond which returns diminish, ultimately reaching optimal performance at 18 views as illustrated in Figure 10. The qualitative results, visualized in Figure 11, validate our multi-view fusion approach while demonstrating the existence of a performance plateau beyond a certain viewpoint threshold.

7. Details on Optimization Strategies

Baseline. Our baseline implementation uses GTA [43] as the feature extraction backbone, achieving 3.27 FPS while maintaining high reconstruction quality. This represents the unmodified network performing full reconstruction at each frame with uniform sampling across the entire volume.

Coordinate Mapping. Introducing coordinate mapping between canonical and posed space initially decreases performance to 2.14 FPS due to the overhead of computing transformation matrices and performing coordinate transformations. This establishes the foundation for canonical space inference and enables subsequent optimizations for tempo-

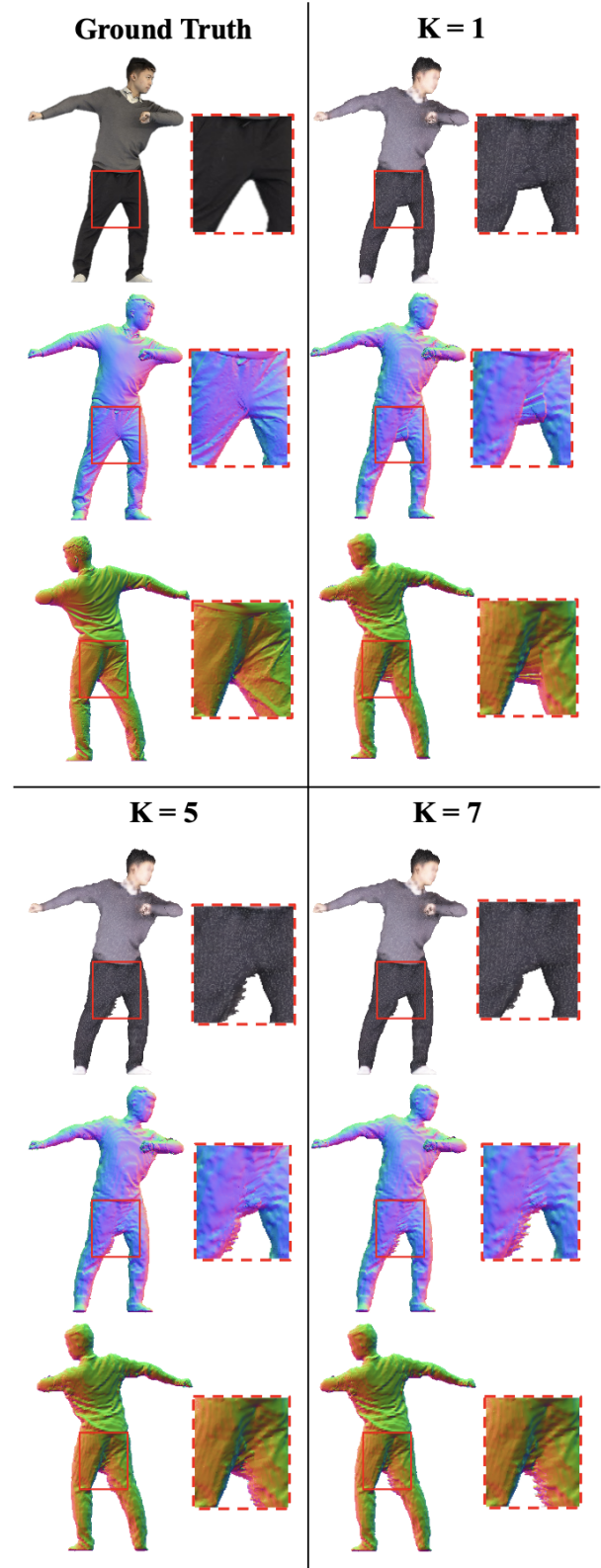


Figure 8. **Qualitative comparison of geometry reconstruction quality.** under different K values in coordinate mapping.

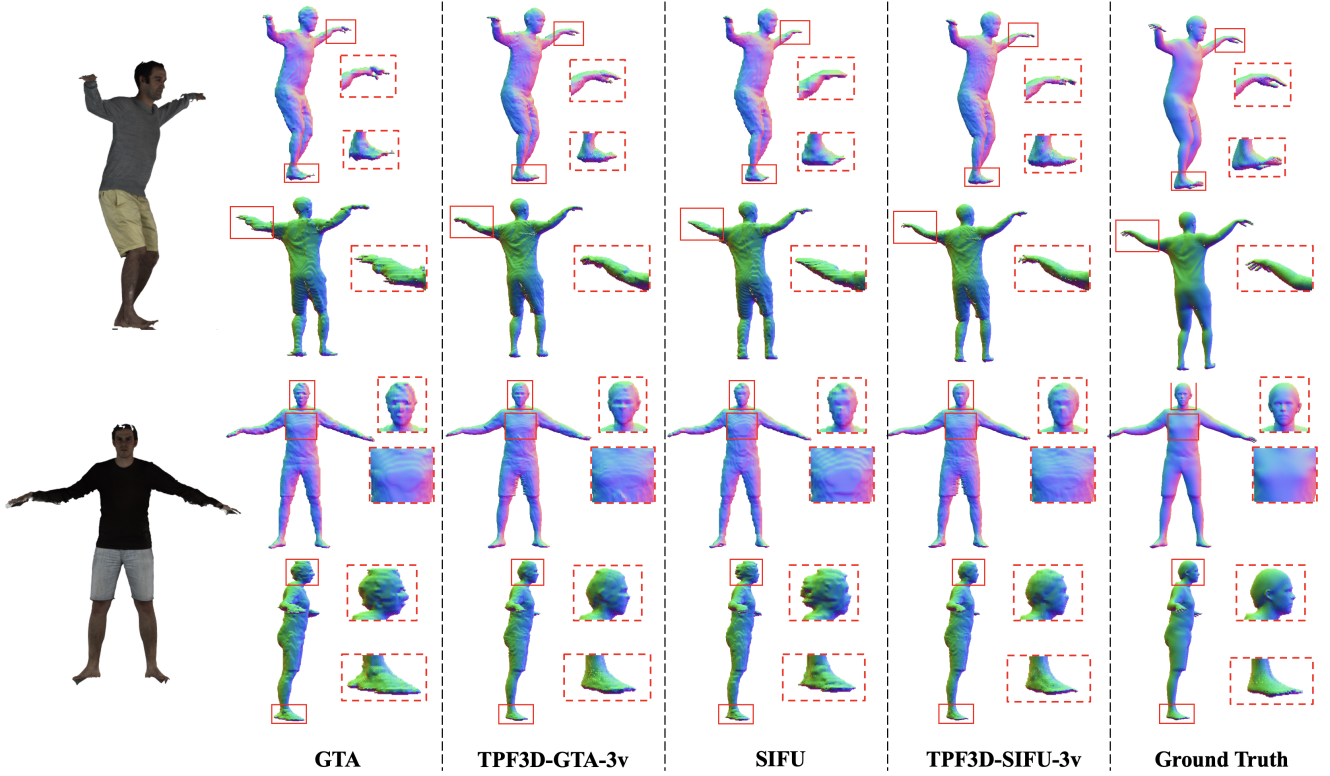


Figure 9. **Qualitative comparison** of geometry reconstruction quality with state-of-the-art methods. Purple: test view, green: novel view.

ral propagation, while temporarily reducing the speed and quality.

Linear Layer. We observe that the query networks \mathcal{G}_s and \mathcal{G}_c contain many 1D convolutional layers with 1×1 filter size, which behave identically to linear layers. Replacing these with actual linear layers increases speed to 2.67 FPS with minimal quality decrease due to implementation differences between linear and convolutional layers in PyTorch.

Visibility-Guided and Surface-Adjacent Sampling. Using visibility-guided sampling alone decreases speed to 1.91 FPS as the number of sampled coordinates remains similar to that of coarse-to-fine inference. However, combining both visibility-guided and surface-adjacent sampling significantly reduces the coordinate count, increasing performance to 4.50 FPS while maintaining reconstruction accuracy comparable to coordinate mapping.

Limited Sampling Points. We further optimize by imposing a strict limit on sampling points ($n < 2^{10}$). This limit is enforced after the two sampling strategies to ensure points are concentrated in dynamically changing regions. As shown in Table 3, this improves speed to 5.84 FPS without sacrificing quality.

TorchScript. The final optimization employs TorchScript compilation to eliminate Python overhead in key computa-

tional operations, achieving $3.77\times$ speedup over baseline (maximum of 12.30 FPS over 3.27 FPS). This optimization focuses on execution efficiency rather than algorithmic modifications, maintaining reconstruction quality with minimal degradation.

8. More results

We provide additional evaluation results to demonstrate our method’s reconstruction capabilities across different scenarios. In Figure 9, we present detailed comparisons with state-of-the-art methods, highlighting the regions with significant differences. Our method (TPF3D-GTA-3v and TPF3D-SIFU-3v) shows improved geometry reconstruction compared to GTA and SIFU baselines. In particular, our approach better preserves fine details in challenging regions such as hands, feet, and head, as shown in the zoomed-in patches. When compared against the ground truth, our reconstructions demonstrate more accurate body proportions and pose estimation, while maintaining geometric details across both test (purple) and novel (green) viewpoints. Figure 12 showcases comprehensive results on the THuman2.0 [40] dataset, displaying reconstructions from three different angles (0° , 120° , 240°).

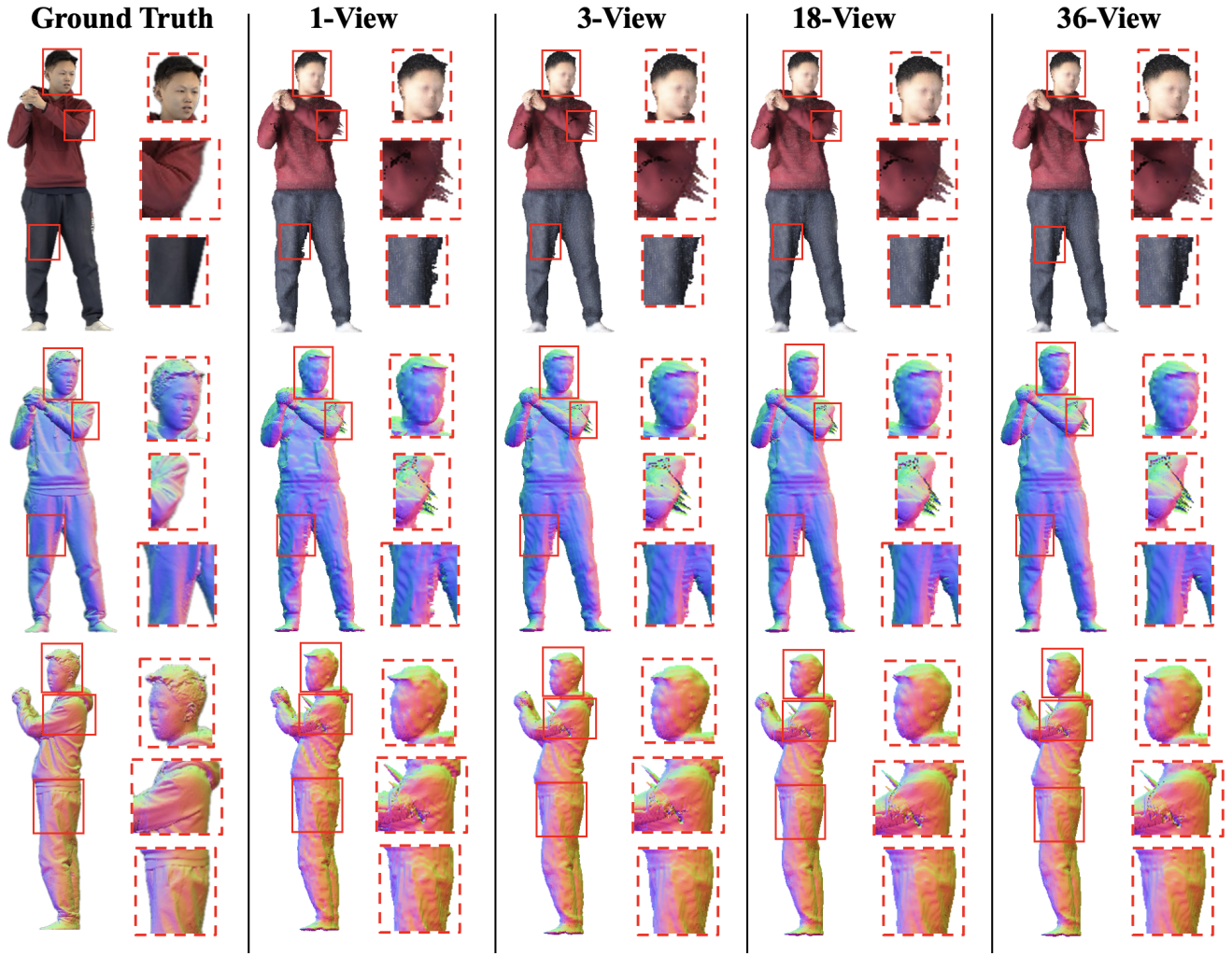


Figure 10. Qualitative comparison of geometry reconstruction quality with varying number of input views

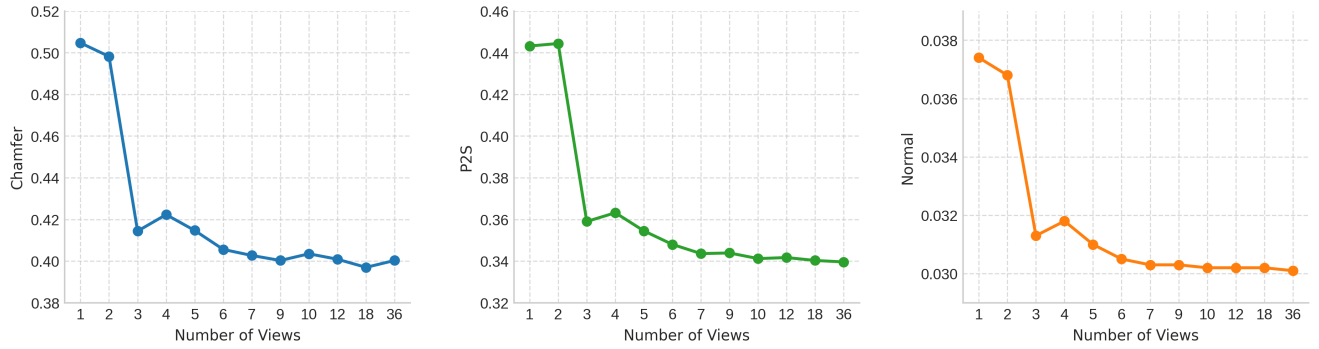


Figure 11. Plotting the results in Table 6 to better visualize the trends in reconstruction quality with respect to number of input views.



Figure 12. **Qualitative results** on the THuman2.0 [40] dataset. (a), (b), and (c) represent 0° , 120° , and 240° test views, respectively. The leftmost column shows the input images, and the rightmost column displays the rendered results on the test view. The purple mesh represents the test view results, while the green mesh corresponds to the novel view results.