FISHuman: Fine-grained Single-image 3D Human Reconstruction via Multi-view 4D Remeshing

Supplementary Material

In our supplementary material, we provide:

- Details of 3D-aware dual-stream video generation.
- Details of dynamic 3D human carving.
- Baseline comparison.
- More results.
- Robustness to inaccurate normal input.
- Limitations.

A. Details of 3D-aware Dual-stream Video Generation

We implement our 3D-aware dual-stream video generator based on the Wan2.1-I2V-14B-720P model [58]. We employ LoRA [16] with a rank of 64 and a scaling factor of 1.0 for 3D-aware cross-modal finetuning. We use Sapiens [24] to predict the normal input based on the reference RGB image. The two-stage training of the video generator takes approximately 5 days in total on 2 A100 GPUs: 4 days for stage 1 and 1 day for stage 2. At inference, the model generates RGB and normal sequences both with 25 frames at the resolution of 1280×720.

B. Details of Dynamic 3D Human Carving

B.1. Details of 4D Remeshing

Given multi-view normal frames with pixel-level inconsistencies, our proposed 4D remeshing module establishes vertex-level correspondence across different viewpoints, by decoupling the learning of the static canonical mesh and view-dependent vertex deformations. Continuous remeshing [40] is adopted to optimize the canonical mesh for its adaptive density control and explicit topology correction, yielding fine-grained details and smooth manifold topology essential for high-quality assets. The positions of the canonical mesh vertices are updated by the optimizer proposed in [40] with a learning rate of 0.3, while the dynamic deformation field is optimized by the Adam optimizer [25] with a learning rate decaying from 1.6×10^{-4} to 1.6×10^{-6}. The normal map $\hat{\mathcal{N}}_i$ and object mask $\hat{\mathcal{S}}_i$ under each view v_i are rendered at the resolution of 1280×720 to compute the reconstruction loss L_{rec}. The ARAP loss L_{arap} during 4D remeshing is calculated by randomly sampling 1K vertices from a pair of deformed meshes corresponding to two different viewpoints.

B.2. Details of Unified UV Representation

The topological consistency of the deformed meshes enables the learning of a unified UV representation to effectively integrate appearance attributes across all perspectives. We initialize the texture map $\mathcal{T}$ at the resolution of 1024^2. It is optimized by the Adam optimizer [25] with a learning rate of 0.01. The view-specific weights w_i are set to 1.0 for the front and back views, and 0.2 for other viewpoints. The optimized UV map undergoes texture enhancement using a pretrained portrait enhancing model [72] that raises the resolution to 2048^2 and improves appearance details.

C. Baseline Comparison

C.1. Baselines

We compare our method with four single-image 3D human reconstruction baselines: SIFU [86], SiTH [12], Human3Diffusion [71], and PSHuman [29], as well as state-of-the-art methods on single-view 3D avatar and object generation: StdGen [11] and Hunyuan3D 2.0 [87]. For Human3Diffusion, which utilizes 3D Gaussian Splatting [23] as the underlying representation, we render the reconstructed 3DGS for appearance evaluation. For geometric comparison on Human3Diffusion, we perform TSDF extraction [80] following the provided scripts to extract explicit meshes. For other mesh-based methods, we follow the default protocol to generate textured meshes for both geometric and appearance comparisons.

C.2. Metric Calculation

We provide detailed implementation for quantitative comparisons with SOTA methods. To ensure a fair and accurate geometric evaluation against the ground-truth meshes, we follow the standard practice of normalizing the reconstructed meshes of all methods to a unit bounding box. This preprocessing step eliminates metric distortions caused by scale and translation differences. We sample 100K surface points to calculate the Chamfer Distance (CD), Point-to-Surface (P2S) distance, and Normal Consistency (NC). For appearance evaluation, we render 25 views evenly around the reconstructed avatars and ground-truth scans for each method at the resolution of 1024^2.

C.3. Comparison with LHM

To further validate the reconstruction quality of our method, we compare it with LHM [43], which represents the state-of-the-art 3DGS-based native 3D human reconstruction approach. Since LHM does not support explicit mesh extraction, we provide a qualitative comparison of the appearance quality in Fig. 9. LHM exhibits moderate reconstruction

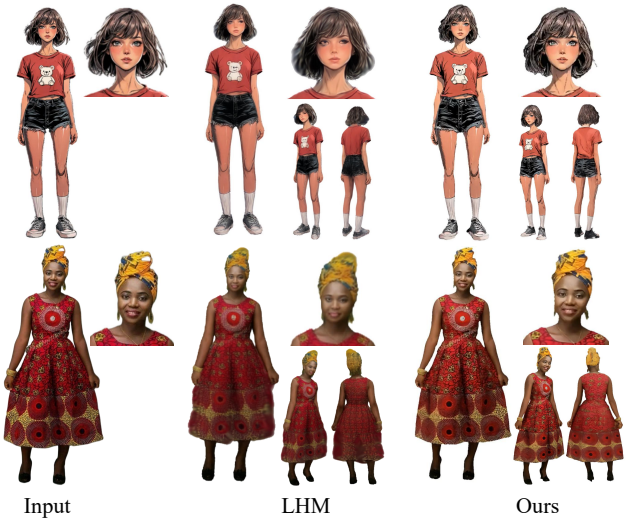

Figure 9. Qualitative comparison with LHM.

fidelity for anime-style characters and complex clothing, while our method enables faithful restoration of the reference images, achieving high-fidelity 3D human reconstruction.

C.4. Comparison with Versatile 3D Object Generation Methods

To contextualize the reconstruction quality of our human-specific approach, we conduct qualitative comparisons against state-of-the-art versatile 3D object generation methods: TripoSG [31], Meshy-4 [55], and Hunyuan3D 2.0 [87]. As demonstrated in Fig. 13, our method achieves comparable geometric and textural details to the general-purpose approaches, while showcasing higher fidelity in human identity preservation and reference image restoration.

D. More Results

D.1. Arbitrary-pose 3D Human Reconstruction

We present additional arbitrary-pose 3D human reconstruction results in Fig. 14, Fig. 15, and Fig. 16. Our method demonstrates strong capabilities in handling diverse real-world cases, including loose clothing, complex accessories, and challenging poses. The generated results exhibit plausible novel views, fine-grained surface details, and photorealistic appearances comparable to real-world individuals.

D.2. Canonical-pose 3D Human Reconstruction

Given an input image of a person in an arbitrary pose, we first achieve human pose standardization using a pre-trained reposing model [60]. This process takes as input the reference image and a pose sequence interpolated from the original pose to a standard A-pose, and outputs the corresponding canonical-pose image of the avatar. Using this preprocessed image as input, we then reconstruct a canonical-pose 3D human with our proposed pipeline. We present additional reconstruction results in Fig. 17, Fig. 18, and Fig. 19. The resulting avatars support direct 3D asset export and can be rigged and animated using standard pipelines (e.g., Mixamo [22]), as shown in Fig. 10.

Figure 10. Novel pose animation of the reconstructed avatars.

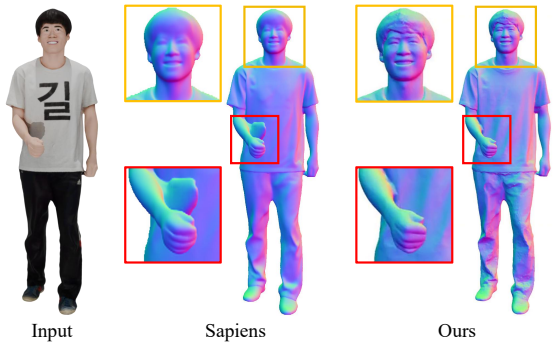

Figure 11. Demonstration of the model's robustness to inaccurate normal input.

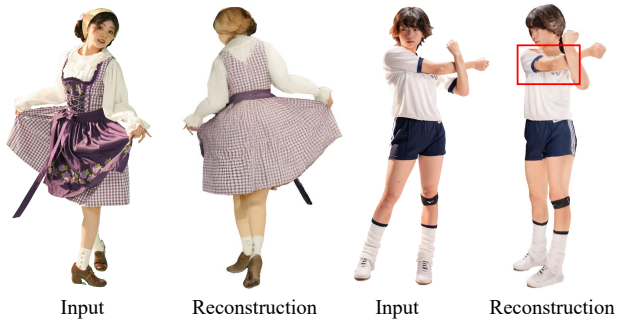

Figure 12. Limitations on cases with complicated patterns or severe self-occlusions.

E. Robustness to Inaccurate Normal Input

The 3D-aware dual-stream video generator takes as input the RGB reference image and its estimated normal map as conditioning signals. However, the normal map predicted by Sapiens [24] may occasionally contain inaccuracies around ambiguous surfaces, such as regions with complex shading or occlusion. As shown in Fig. 11, our video generator demonstrates strong robustness to such imperfect normal inputs, effectively correcting the estimation errors. Furthermore, it also enhances the detail quality of normal surfaces, leading to finer geometric reconstruction.

F. Limitations

F.1. Complicated Patterns

Although our dual-stream video generation model synthesizes plausible and realistic novel views in most cases, it faces challenges with intricate clothing patterns, such as dense and large-area plaids, resulting in texture blurring. As shown in Fig. 12, the degraded RGB guidance subsequently propagates to the 3D reconstruction, leading to visual artifacts in novel-view renderings.

F.2. Severe Self-occluded Regions

While our cross-modal alignment strategy ensures strict correspondence between the RGB and normal guidance at the pixel level, the dynamic reconstruction process may still introduce subtle misalignment between the optimized mesh geometry and multi-view RGB frames. This can lead to incorrect color mapping in regions with severe self-occlusions, as shown in Fig. 12. Further refinement of the optimized UV map with reference-image awareness could potentially mitigate this issue.

F.3. Complex lighting

Our video generator is trained on renderings with uniform lighting conditions, which may exhibit artifacts when faced with real-world images under uneven illumination and overexposure. Disentangling intrinsic albedo from complex lighting conditions during generation is a possible approach to enable production-ready 3D asset creation with accurate material properties.

F.4. Efficiency

While our method generates production-ready 3D assets, it requires approximately 7–8 minutes per reconstruction on an A6000 GPU, which is slower than recent feed-forward methods. This limits its applicability in interactive or real-time scenarios.

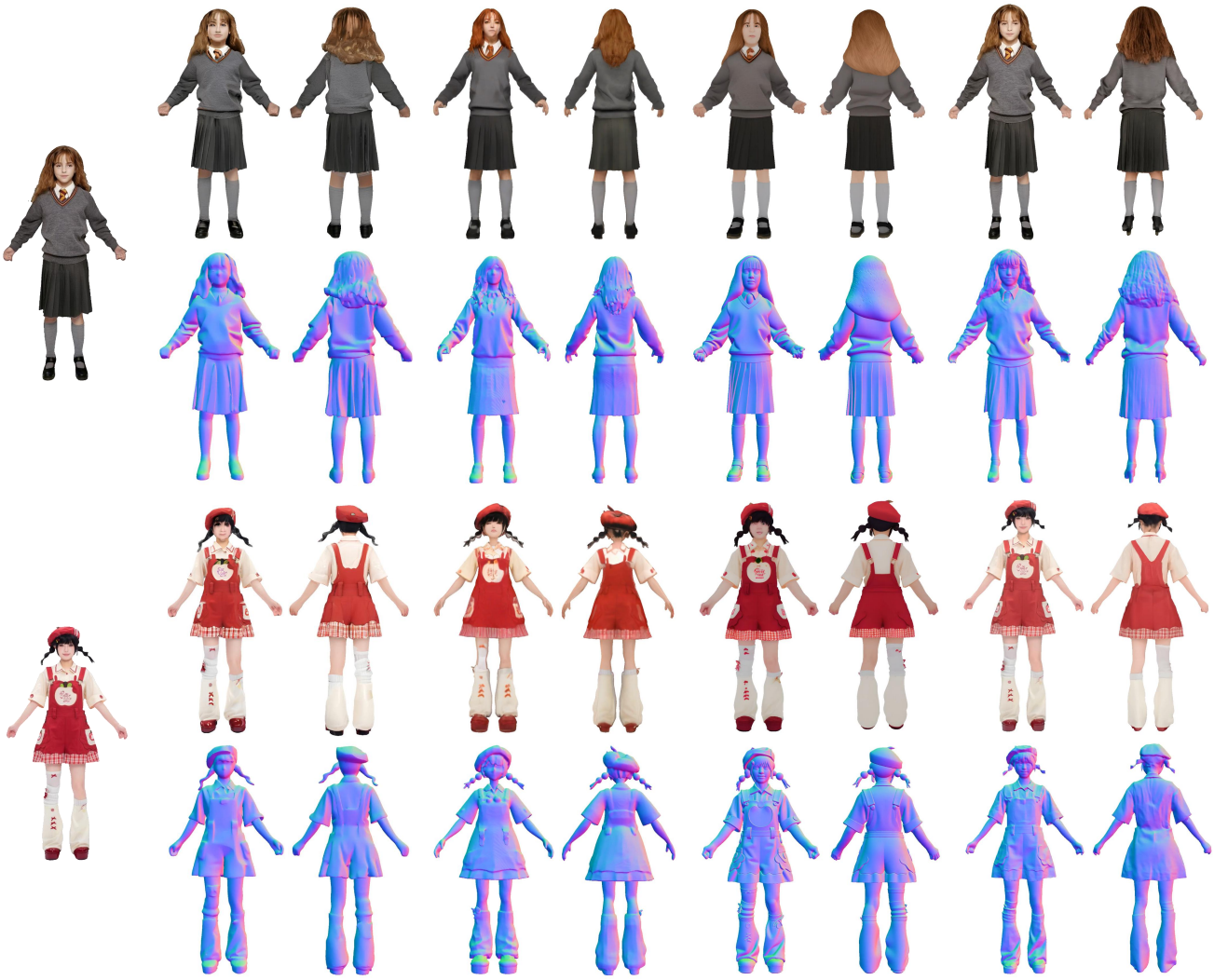

Figure 13. Qualitative comparison with versatile 3D object generation methods.

Figure 14. More results on arbitrary-pose 3D human reconstruction.

Figure 15. More results on arbitrary-pose 3D human reconstruction.

Figure 16. More results on arbitrary-pose 3D human reconstruction.

Figure 17. More results on canonical-pose 3D human reconstruction.

Figure 18. More results on canonical-pose 3D human reconstruction.

Figure 19. More results on canonical-pose 3D human reconstruction.