PSHuman: Photorealistic Single-image 3D Human Reconstruction using Cross-Scale Multiview Diffusion and Explicit Remeshing

Supplementary Material

6. Discussions about face-body cross-scale diffusion

Difficulty in implementing dependent forward process. In the dependent forward process $q(x_t^B|x_{t-1}^B, x_{t-1}^F)$, we know that the face region of x^B corresponds to x^F. Since we have defined $p(x_t^F|x_{t-1}^F)$ by adding noises to x_{t-1}^F, it is natural to get x_t^B by replacing the pixel values in the face region of x_t^B with x_t^F and just adding noises to the remaining image regions of x_{t-1}^B. However, since we adopt a latent diffusion model (Stable Diffusion) Rombach et al. [33] here, the pixels of tensors in the latent spaces are not independent of each other so the replacing operation is not valid here. This brings difficulty in separating the face regions in the latent space to explicitly implement the dependent forward process for adding noises.

Rationale of approximated forward process. Our rationale for adding noises to the face and the body separately is that the process is similar to multiview diffusion. We can regard the face image and the body image as just two images captured by cameras with different camera positions and focal lengths. In this case, the body-face cross-scale diffusion is a special case of multiview diffusion. In a multiview diffusion, we add noises to multiview images separately so that we can also add noises to the body image and face image separately but consider the dependence in the reverse process.

7. Implementation Details

Preprocessing. Our training datasets include scans from THuman2.1 and CustomHumans. For each human model, and the corresponding SMPL-X model, we render 8 color and normal images with alpha channel around the yaw axis, with a $45°$ interval and a resolution of 768×768. Due to the random face-forward direction, we employ insightface Deng et al. [5] for face detection, utilizing only viewpoints containing clear facial characteristics for training.

Choice of generated views. As mentioned in the main paper, PSHuman generates 6 color and normal images from front, front-right, right, back, left, and front-left views for the trade-off between effectiveness and training workload. To guarantee the generation alignment, we horizontally flip the left and back views during training. In Fig. 12, we present the results reconstructed using only two-view (front and back) or four-view (front, right, back, left) normal maps. Since there is a lack of depth in information, optimizing geometry with fewer views leads to severe artifacts, such as incomplete or unnatural human structures. In contrast, it is evident that the artifacts are reduced when using six views.

Table 6. Inference time of the reconstruction module.

Pipeline	Pre-processing	Diffusion	Geo. Recon.	Appearance Fusion
Time / s	7.2	17.6	23.3	6.0

Table 7. User study w.r.t reconstruction quality and novel-view consistency.

Method	PIFuHD	PaMIR	ECON	GTA	SiTH	Ours
Geometry Quality	1.55	1.96	3.72	2.11	2.72	4.71
Appearance Quality	-	1.42	-	2.65	2.82	4.59
Geometry Consistency	1.69	1.76	2.48	2.33	2.79	4.61
Appearance Consistency	-	1.77	-	2.16	2.73	4.68

Diffusion block. As illustrated in Fig.3(b) of the main paper, our diffusion block comprises two branches. The local diffusion inherits from stable diffusion (SD2.1-Unclip) [34], including self attention, cross attention and feed-forward layers, while the global attention contains an additional multi-view attention layer introduced in Era3D [21]. The global attention is conditioned on the local branch via the noise blending layer. We feed the embeddings of text prompt ”a rendering image of 3D human, [V] view, [M] map.” into the denoising blocks via cross attention, where [V] is chosen from ”front”, ”front right”, ”right”, ”back”, ”left”, ”front left”, ”face” and [M] represents ”normal” or ”color”.

Inference details. Given a human image, we first remove the background with rembg [7] and then resize the foreground to 720×720. Finally, we pad it to 768×768 and set the background to white. Due to the alignment between of processed input image and the generated front color image, we use the former and other generated images in the following reconstruction.

8. More experiments

Inference time. In Tab. 6, we report the detailed inference time of the whole pipeline, including preprocessing (SMPL-X estimation and SMPL-X image rendering), diffusion, geometry reconstruction (SMPL-X initialization and remeshing) and appearance fusion.

User study Given the limitations of quantitative metrics in assessing the realism and consistency of side and back views reconstructed from single-view input, we conducted a comprehensive user study to evaluate the geome-

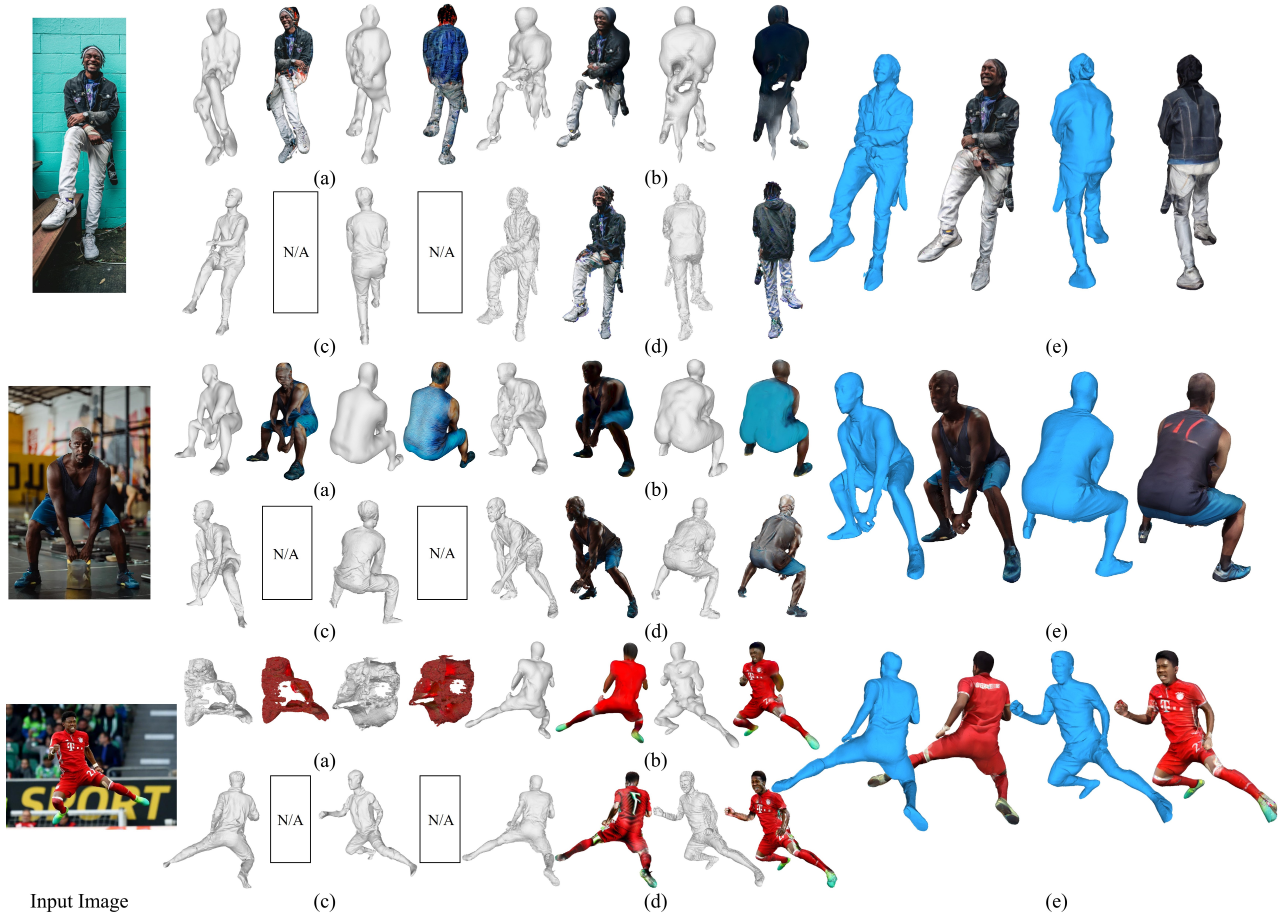

Figure 11. Qualitative comparison with optimization-based methods. We demonstrate the results of **(a)** Magic123, **(b)** Dreamgaussian, **(c)** Chupa, **(d)** TeCH and **(e)** Ours.

try and appearance quality of five SOTA methods. Specifically, we collect 20 in-the-wild samples and 20 cases from SHHQ fashion dataset for evaluation. Following HumanNorm [14], we invite 20 volunteers to evaluate the color and normal video rendered from the reconstructed 3D humans. Participants were instructed to score each model on a 5-point scale (1 being the worst and 5 being the best) across four key dimensions:

- To what extent does the human model exhibit the best geometry quality?
- To what extent does the human model exhibit the best appearance quality?
- To what extent does the novel view's geometry of the human body align with the reference image?
- To what extent does the novel view's appearance of the human body align with the reference image?

For methods that do not produce texture (PIFuHD and ECON), we only compare the geometry quality and consistency. The results in Tab. 7 indicate that our method represents a significant advancement against SOTA methods, offering superior performance in both geometry and appearance reconstruction, as well as consistency across novel viewpoints.

Comparison with optimization-based methods. To assess the efficacy of our approach relative to optimization-based methods, we conducted a comparative analysis of PSHuman against several SDS-based techniques, Magic123, Dreamgaussian, Chupa, and TeCH. Following SiTH, we adopt the pose and text prompt generated by [20] as condition inputs due to the lack of direct image input support in Chupa. As illustrated in Fig. 11, Magic123 and Dreamgaussian exhibit significant limitations, primarily manifesting as incomplete human body reconstructions and implausible free-view textures. The reliance on text descriptions for conditioning proves insufficient for fine-grained control, resulting in geometries that deviate substantially from the reference inputs. TeCH, a method specifically designed for human reconstruction from a single image, while capable of producing complete human shapes, struggles with severe noise in geometric details and over-

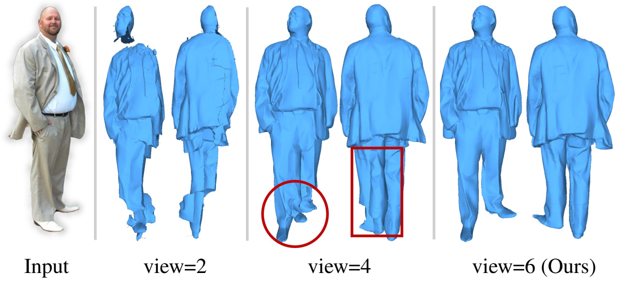

Figure 12. Ablation of view number. Since normal maps lack depth information, optimizing geometry by only two or four views leads to an incomplete or unnatural human structure.

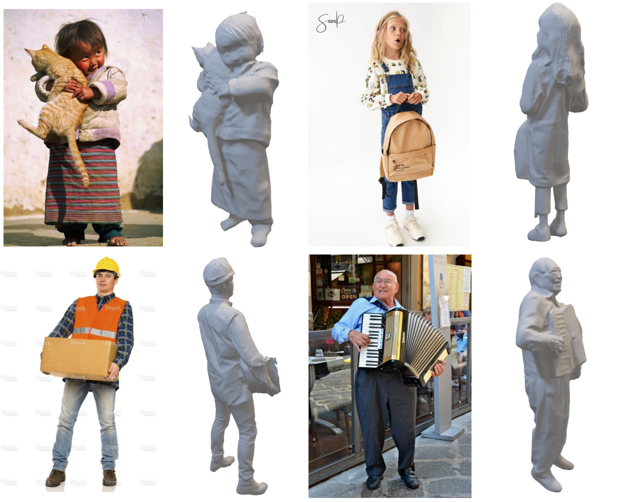

Figure 13. Reconstruction quality on object-occluded images.

saturated textures. These artifacts are characteristic challenges inherent to SDS-based methodologies. In contrast, PSHuman demonstrates superior performance by directly fusing multi-view 2D images in 3D space, enabling the preservation of geometry details at the pixel level while circumventing unrealistic texture. Note that TeCH requires $\sim$6 hours for optimization, PSHuman generates high-quality textured meshes within merely 1 minute. We refer readers to Fig. 20 and Fig. 21 for more results generated by PSHuman.

Capability of handling occlusion. We present the generated normal maps (back, left, and right views) and corresponding meshes of in-the-wild samples with various self-occlusion, as demonstrated in Fig. 18. To further illustrate the robustness of our approach, we also include examples of object-occluded scenarios in Fig. 13. The results show that our diffusion model can infer the correct human structure under both self-occlusions and object occlusions, enabling the reconstruction of high-quality 3D meshes even under such challenging conditions.

Robustness to SMPL-X estimation. The SMPL-X serves as a coarse anatomy guide, only required to be reasonably overlayed with the human body. Thus, our method could handle estimation error (Fig. 14) to some extent and generalize to children or the elder in Fig. 15.

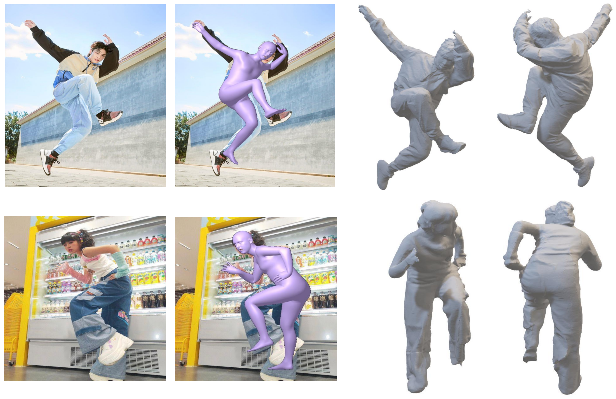

Figure 14. Robustness to SMPL-X estimation errors.

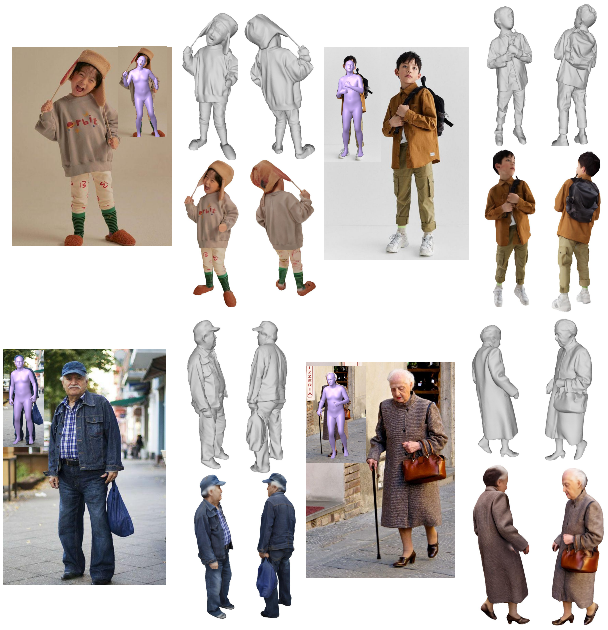

Figure 15. Performance with out-of-distribution pose estimation, like children and the elder.

Robustness to lighting. By incorporating varying lighting conditions using HDR maps from Poly Haven during training, our model demonstrates robustness to lighting variations, as illustrated in Fig. 16.

Comparisons of face normal estimation. As shown in Fig. 17, our local face diffusion model generates facial normal images with significantly enhanced fine-grained details compared to ECON [43] and SAPEIN-2B [18].

Generalization on anime characters. Our model, trained with only realistic human scans, exhibits excellent generalization on anime or hand-drawn style character images, as shown in Fig. 19. This is because our method is adapted from the Stable Diffusion [34] model, which has been trained on images of various styles. Thus, our method main-

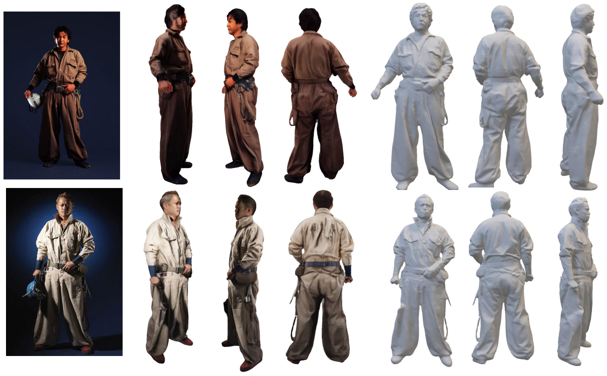

Figure 16. Robustness to shading and strong light.

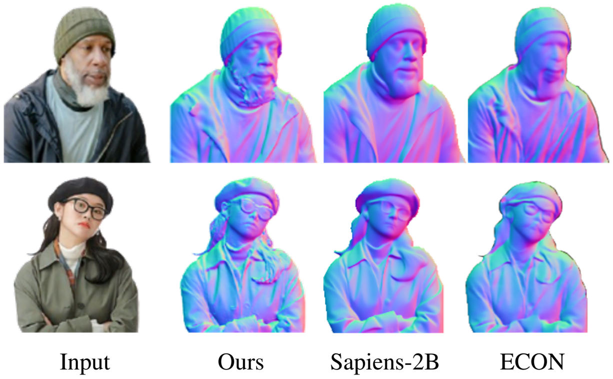

Figure 17. Comparisons of face normal estimation.

tains the ability to generalize images of different domains.

9. Ethics statement

While PSHuman aims to provide users with an advanced tool for single-image full-body 3D human model reconstruction, we acknowledge the potential for misuse, particularly in creating deceptive content. This ethical concern extends beyond our specific method to the broader field of generative modeling. As researchers and developers in 3D reconstruction and generative AI, we have a responsibility to continually address these ethical implications. We encourage ongoing dialogue and the development of safeguards to mitigate potential harm while advancing the technology responsibly. Users of PSHuman and similar tools should be aware of these ethical considerations and use the technology in accordance with applicable laws and ethical guidelines.

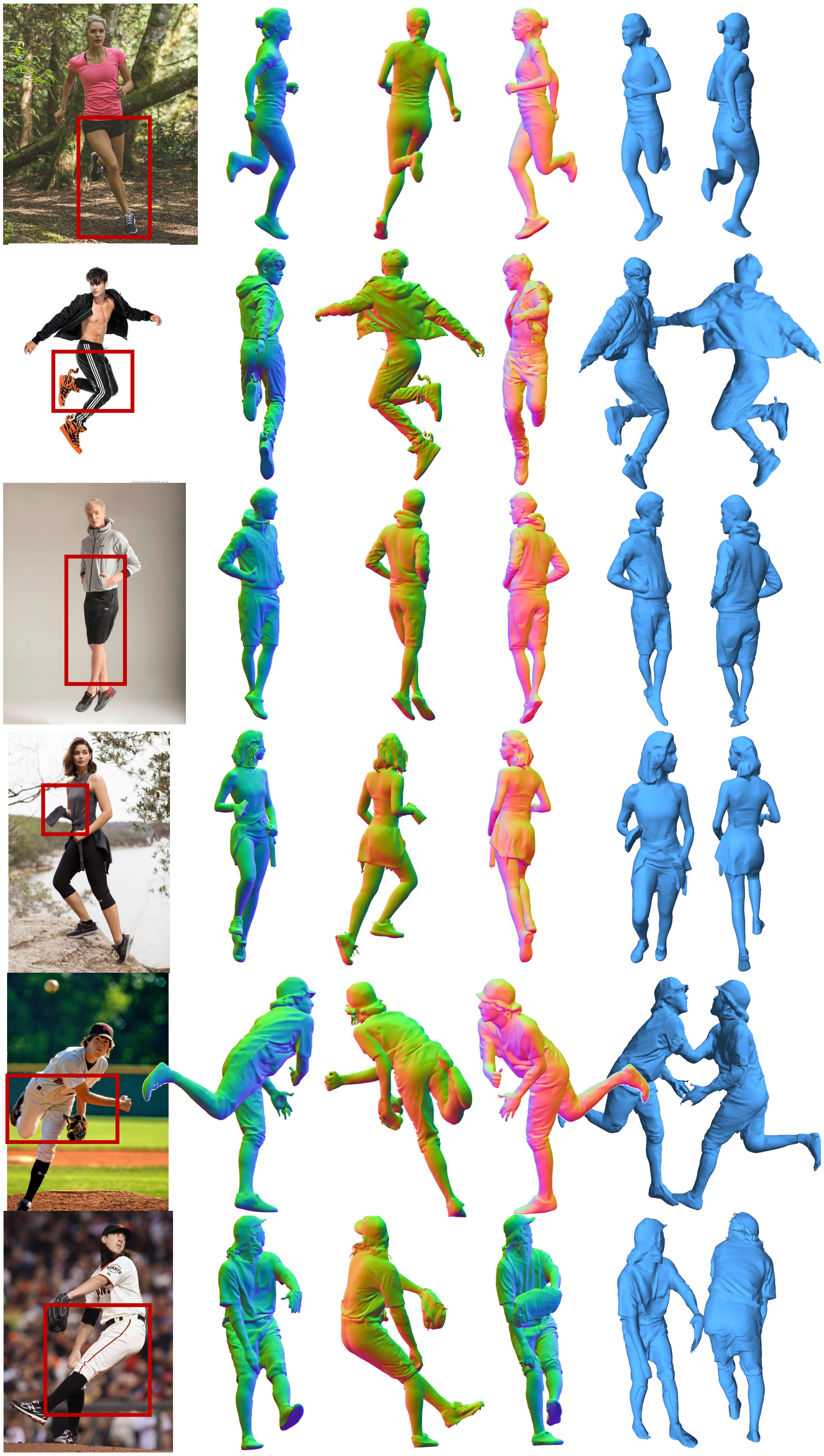

Figure 18. Reconstruction quality on **self-occluded** images. We present the generated back, left, and right views of normal maps and corresponding meshes.

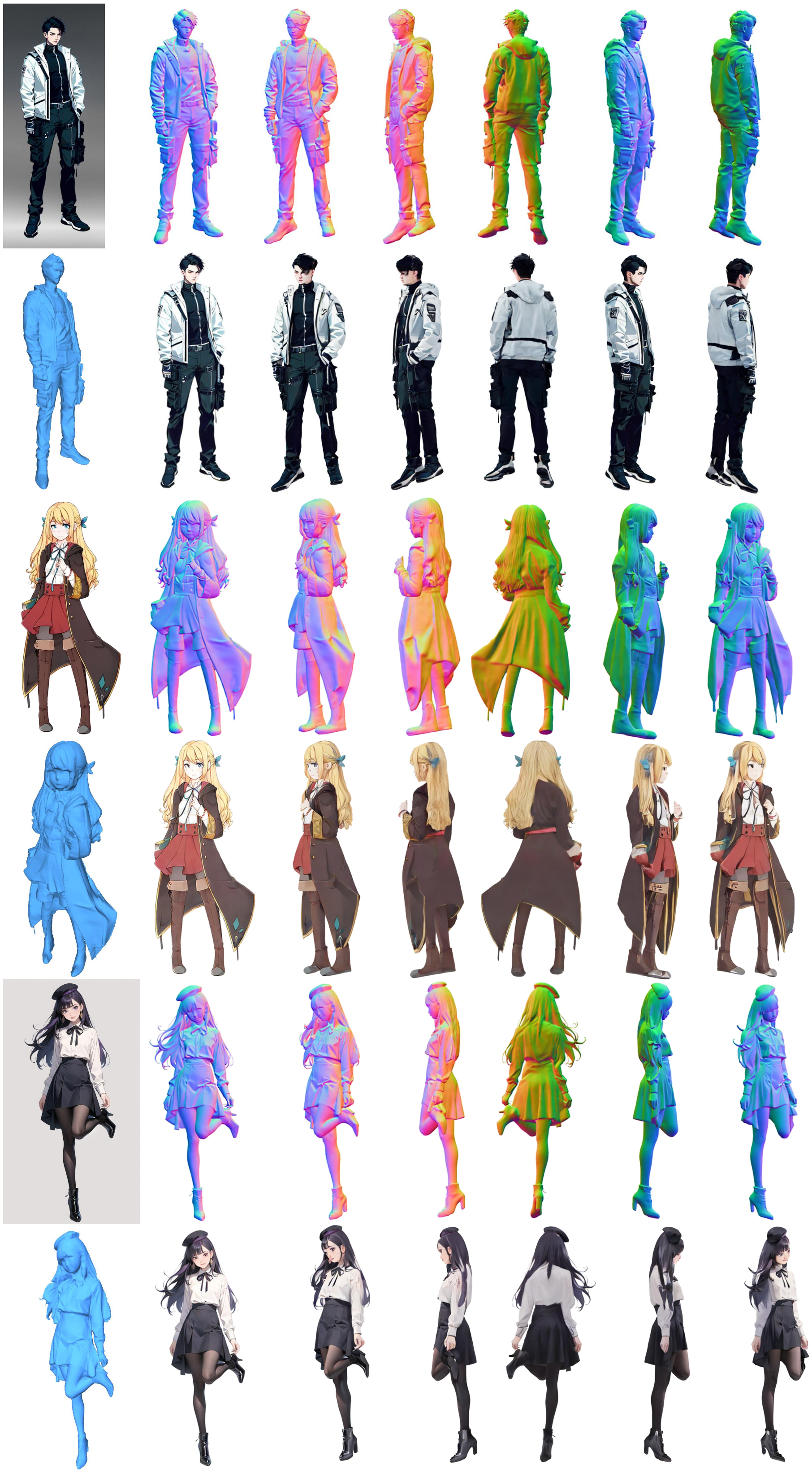

Figure 19. Generalization on anime characters. We present the generated multiview color and normal images and corresponding meshes (in blue).

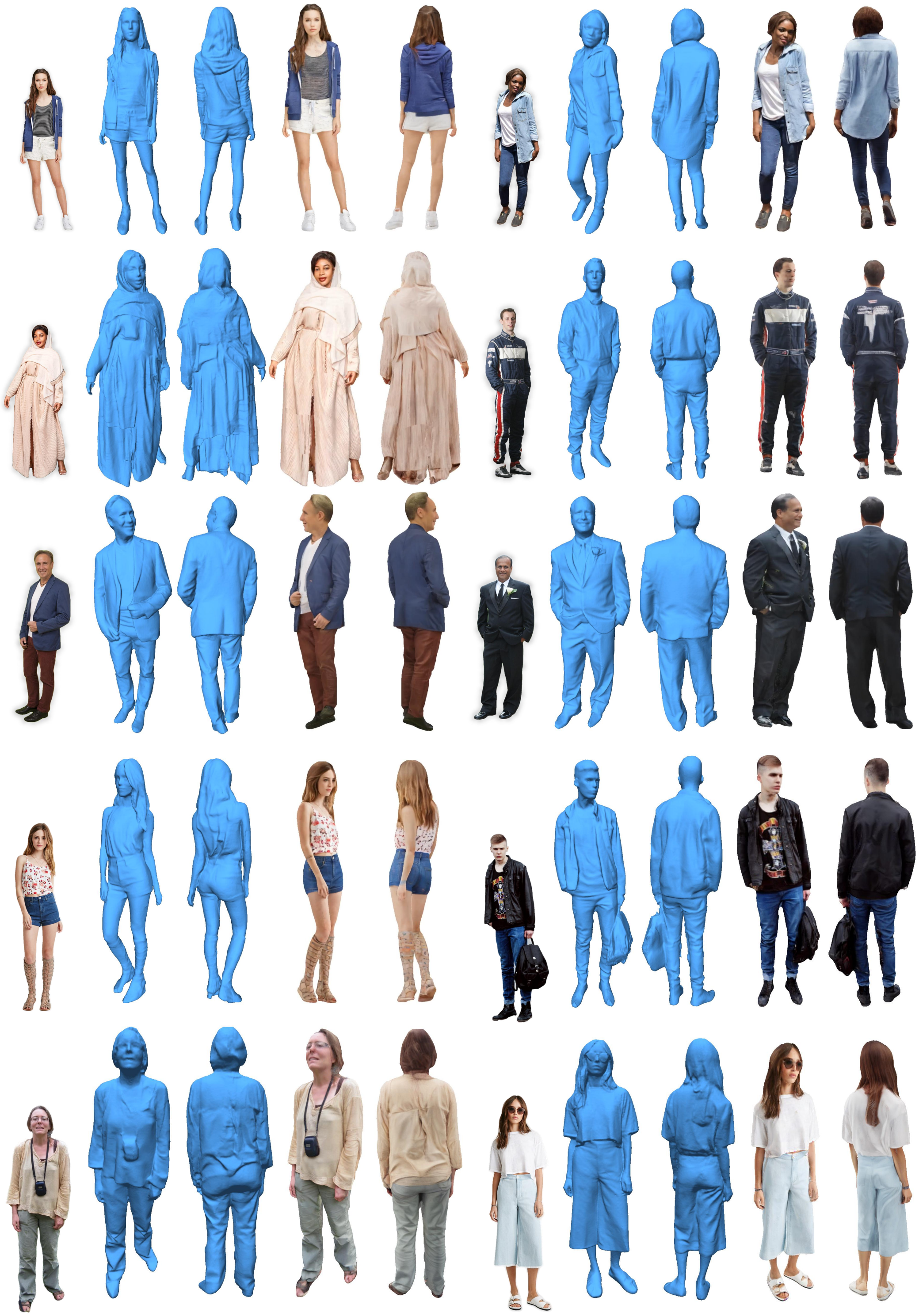

Figure 20. More results on SHHQ dataset.

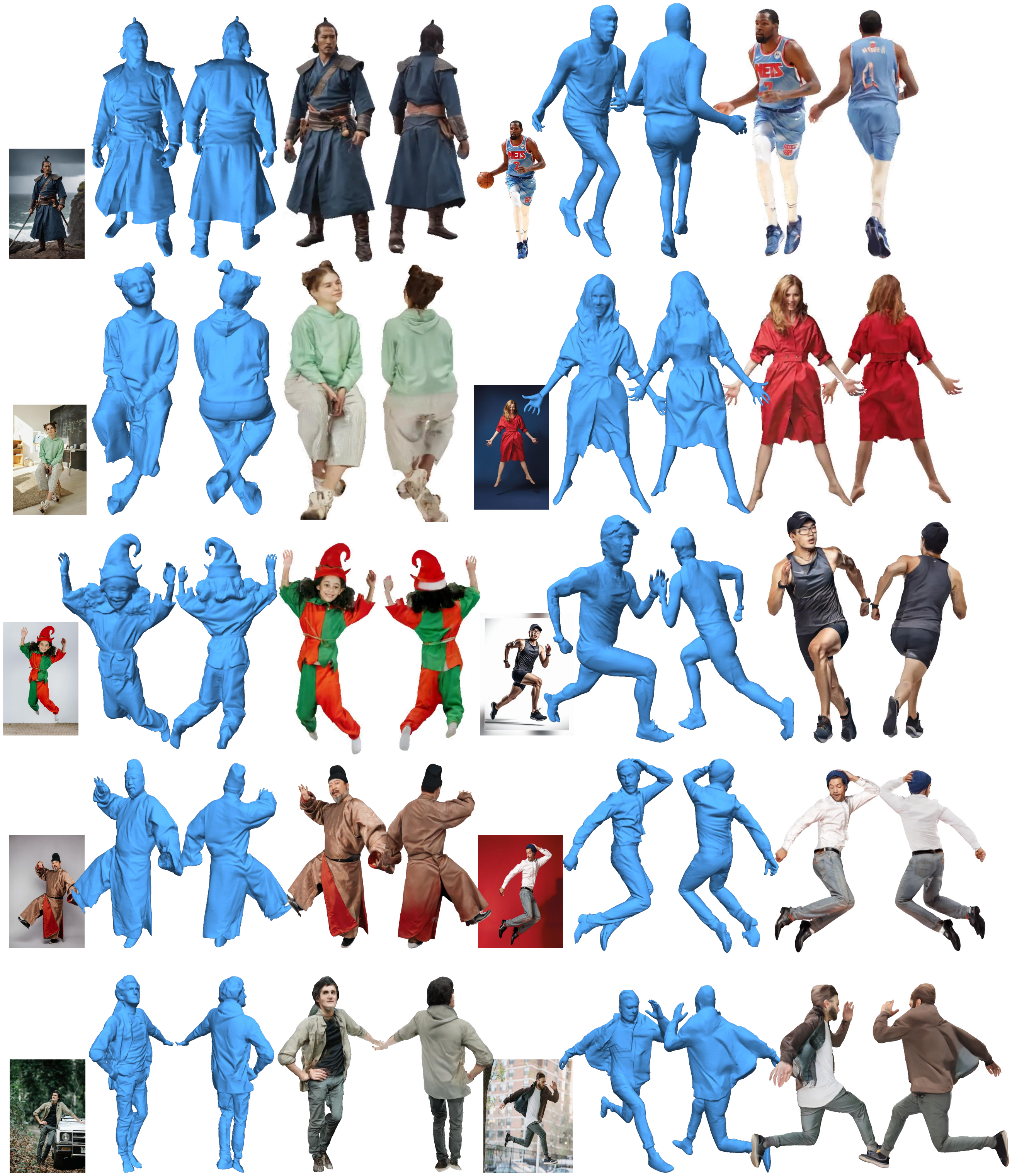

Figure 21. More results on in-the-wild data.