HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images – Supplementary Material

Xihe Yang[1,2,†*] Xingyu Chen[1,✉*] Daiheng Gao[4] Shaohui Wang[1,5,†]
Xiaoguang Han[2,3] Baoyuan Wang[1,✉]
[1]Xiaobing.AI [2] SSE, CUHKSZ [3] FNii, CUHKSZ [4]Freelancer [5]Tsinghua University
https://seanchenxy.github.io/HaveFunWeb/

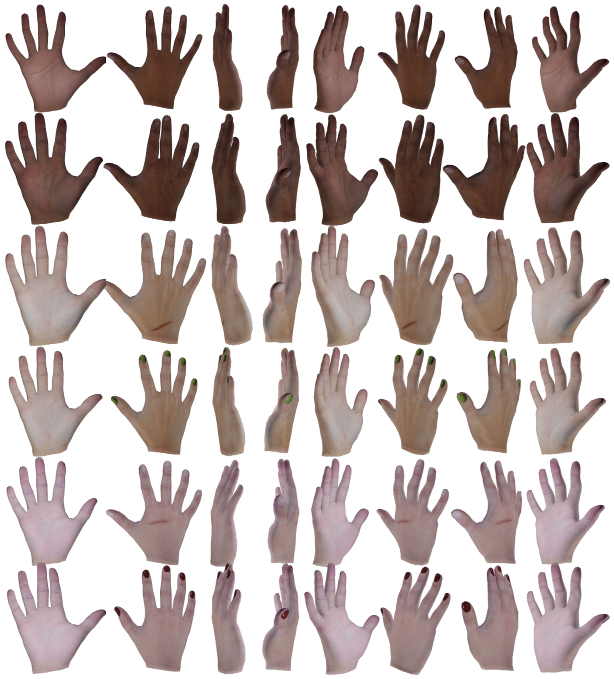

Figure I. Training data for Fig. 6 of the main text. The first N samples in rows are used for the N-shot task.

I. Dataset Details

FS-DART. Our FS-DART is a synthetic dataset based on the DART [3] hand model. We create 100 hand identities with a variety of skin colors and hand shapes. In addition, special hand features such as scars, moles, and nail polish are also included in hand textures. As for hand poses, we capture real hand videos and extract pose parameters with MobRecon [1] for training data creation. As shown in Fig. I, our training samples contain unconstrained casual hand poses. Note that self-occluded poses are not involved so that few-shot data can exhibit sufficient information for the hand reconstruction task.

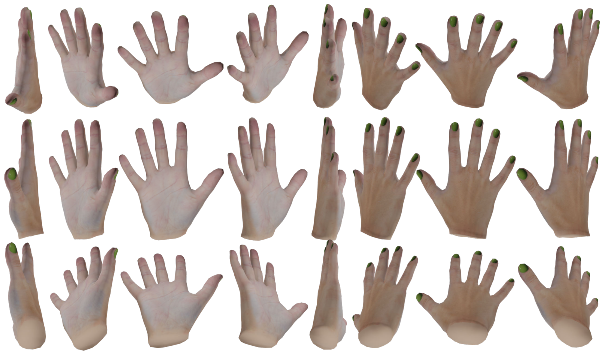

Figure II. Evaluation data of FS-DART from 24 viewpoints.

In terms of evaluation, we assess the effectiveness of our method under the zero hand pose to unveil the performance in shape and texture reconstruction. Referring to Fig. II, the hand in the zero pose is rendered from 24 sphere-distributed viewpoints, and our model can generate corresponding results for metric computation.

FS-XHumans. Our FS-XHumans dataset is built on really captured XHumans [6], which is a 3D scan dataset with 19 actual human identities. For each individual, the XHumans provide 3D scans of motion sequences, including diverse body poses, hand gestures, and facial expressions. Thereby, we select 8 scans from a sequence to produce training data. During data selection, we ensure the diversity of poses and expressions for our training samples, as illustrated in Fig III. Due to the absence of canonical-pose samples, we opt for a scan closely resembling the A-pose to generate testing samples from 24 sphere-distributed view-

[*] Equal contribution; ✉ Corresponding author.
[†] This work was done when Xihe Yang and Shaohui Wang were interns at Xiaobing.AI, led by Xingyu Chen.

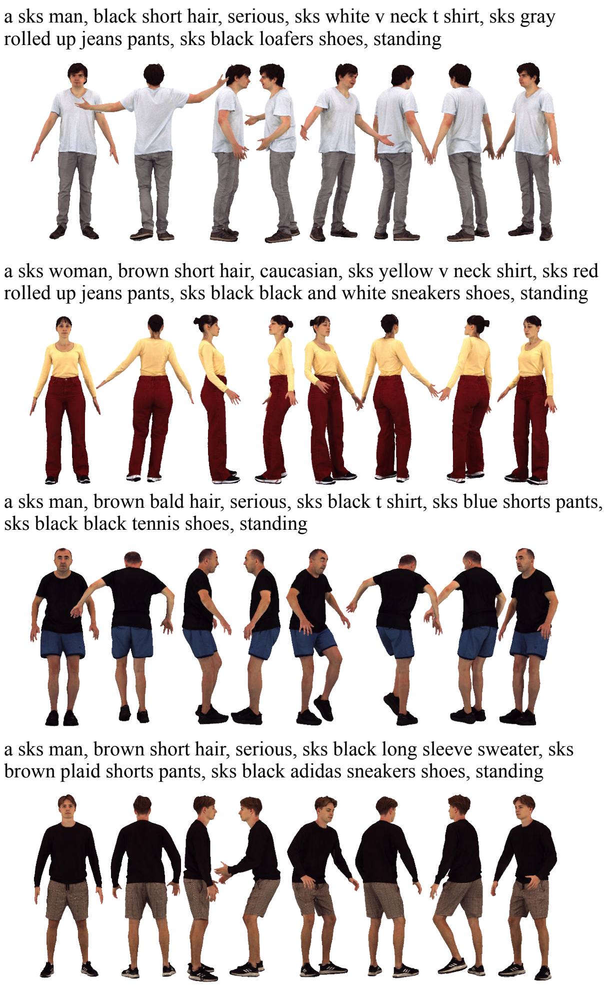
a sks man, black short hair, serious, sks white v neck t shirt, sks gray rolled up jeans pants, sks black loafers shoes, standing
a sks woman, brown short hair, caucasian, sks yellow v neck shirt, sks red rolled up jeans pants, sks black black and white sneakers shoes, standing
a sks man, brown bald hair, serious, sks black t shirt, sks blue shorts pants, sks black black tennis shoes, standing
a sks man, brown short hair, serious, sks black long sleeve sweater, sks brown plaid shorts pants, sks black adidas sneakers shoes, standing

Figure III. Training data for Fig. 7 of the main text. The first N samples in rows are used for the N-shot task. The textural captions are only employed by TeCH.

points for metric computation, as depicted in Fig. IV.

It is worthwhile to note that the training data do not have to strictly follow viewpoints in Figs. I and III. We use this viewpoint configuration as an example because it is an efficient setting for few-shot data acquisition. The viewpoints of arbitrarily captured data can be obtained through parametric geometry estimation [1, 5]. From the perspective of real-world applications, our data setup is reasonable because obtaining data similar to Figs. I and III in practical capture scenarios is straightforward.

II. Implementation Details

Tetrahedral grid. We produce a tetrahedral grid in a 128^3-size cube using 277,410 vertices and 1,524,684 tetrahedra. Positional displacements and an SDF value are attached to vertices, and we explicitly treat them as optimization parameters without resorting to neural networks.

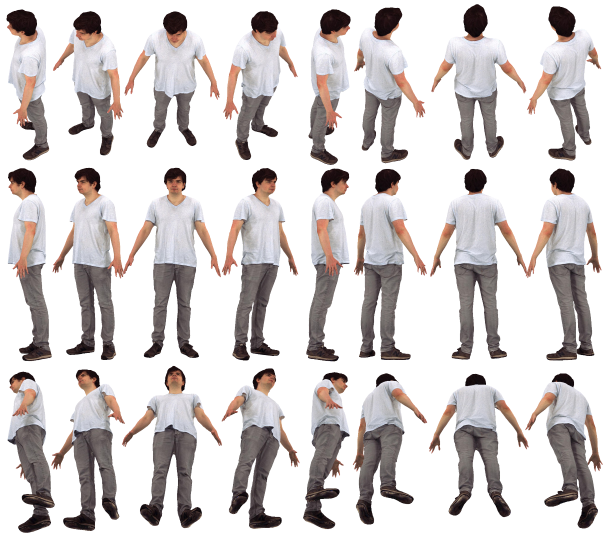

Figure IV. Evaluation data of FS-XHumans from 24 viewpoints.

Texture field. To predict RGB values, we design texture filed $\mathcal{C}$ using a 3-layer MLP network with a hidden dimension of 64 and a hash positional encoding with a maximum resolution of 2048 and 16 resolution levels. Specifically, the triangle mesh $\mathcal{M}$ extracted from DMtet is deformed to match the posed human space aligned with the training images. Each pixel is mapped onto the deformed mesh surface, represented by its barycentric coordinates. Then, we query points $\mathbf{P}_s$ on the canonical triangle mesh $\mathcal{M}$ with the barycentric coordinates, and the rendered image can be obtained with $\hat{\mathbf{I}} = \mathcal{C}(\mathbf{P}_s)$.

Optimization details. Our experiments are conducted on a NVIDIA A100 GPU. The whole framework is trained in an end-to-end manner.

For the body reconstruction task, the optimization comprises 17,000 iterations. The learning rate starts at 0.05 and is decreased by a factor of 0.1 at the 7,500th and 15,000 steps. The optimization process for a human body takes approximately 4 hours.

In terms of the hand reconstruction task, the optimization requires 2,000 iterations with a learning rate of 0.05. The optimization of a hand identity only costs about 10 minutes.

III. Details of Compared Methods

Due to the absence of existing methods designed for few-shot dynamic human reconstruction, we compare HaveFun with a video-based approach and a one-shot static pipeline.

SelfRecon. In contrast to our data configuration, SelfRecon [4] is designed for self-rotated video data. Despite this difference, SelfRecon can perform human reconstruction under our data setup. That is, few-shot unconstrained images used in our work can be treated as key frames of a video. Hence, it is reasonable to compare our approach with SelfRecon. To this end, we acquire officially released implementation codes from https://github.com/jby1993/SelfReconCode and re-implement the part of the dataset for the adaptation of few-shot image input. In addition, we set a batch size of 2 and a training step of 15,000. The training process costs about 12 hours for a human individual. Furthermore, we also train a SelfRecon model following its original data setting. That is, we generate video data consisting of 100 frames, containing uniformly self-rotated body images, as shown in "SelfRecon (100-shot)" in Fig. 7 of the main text. The SelfRecon results are also displayed in our *suppl. video*. As shown, the instability in geometry and texture is evident across different viewpoints due to the employed training samples with highly articulated motion and the intrinsic mechanism of viewpoint-dependent color prediction.

For the hand experiment, we integrate MANO articulation into SelfRecon and adopt the same settings as the body experiment.

TeCH. TeCH is a one-shot human reconstruction method utilizing SDS guidance, similar to the technical pipeline in our HaveFun framework. For comparison, we employ the official implementation from https://github.com/huangyangyi/TeCH. TeCH requires 5 stages to optimize a human avatar, including VQA caption, DreamBooth fine-tuning, geometry optimization, geometry post-processing, and texture optimization. The captions used for text-guided SDS loss are shown in Fig III. In addition, we argue that the stage of geometry post-processing is tricky due to the replacement of the hand shape with the SMPLX hand mesh. That is, the hand is reconstructed using SMPLX rather than TeCH. For a fair experimental setup, we omit the geometry post-processing and jointly optimize the complete geometry and texture. All other settings adhere to the original TeCH report, and it takes approximately 6 hours to generate a human avatar.

As the VQA caption of hands is unexplored, we do not include the comparison of TeCH in the hand task.

IV. More Results

Effects of normal and depth losses. Referring to Table I and Fig. V, normal and depth losses give rise to instructive effects on human avatar reconstruction. Nevertheless, removing depth loss only leads to a minor performance drop. Due to the often inaccurate estimates of monocular depth, depth supervision is optional in real-world applications, and the HaveFun framework can present human avatars without depth labels.

$\mathcal{L}_{normal}$	$\mathcal{L}_{depth}$	PSNR ↑	SSIM ↑	LPIPS ↓
		4-shot FS-XHumans		
✓	✓	**25.64**	**0.9627**	**0.0347**
✓		25.08	0.9601	0.0352
	✓	24.30	0.9581	0.0404
		23.85	0.9575	0.0466

Table I. The effects of normal and depth losses.

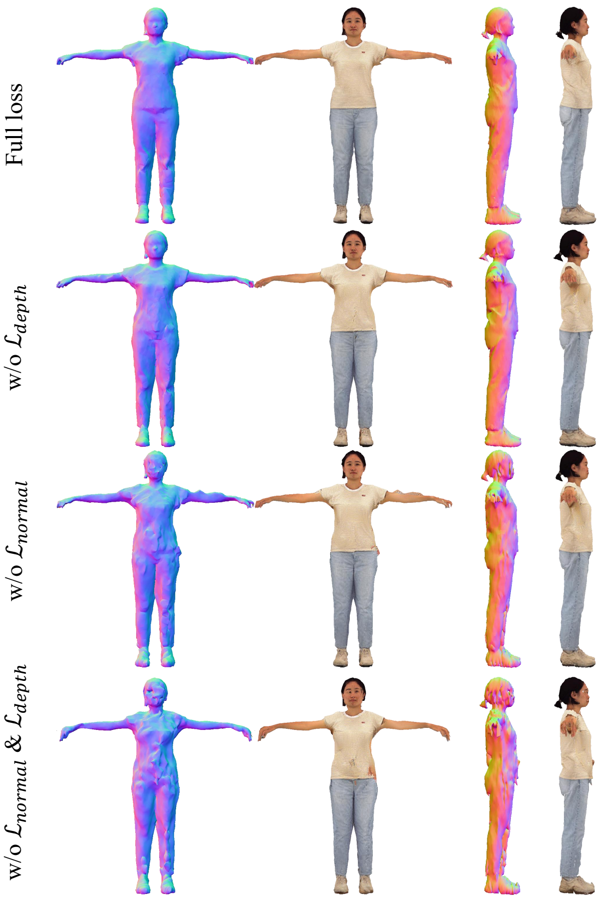
Full loss
w/o $\mathcal{L}_{depth}$
w/o $\mathcal{L}_{normal}$
w/o $\mathcal{L}_{normal}$ & $\mathcal{L}_{depth}$

Figure V. The effects of normal and depth losses.

SDS loss for the 8-shot task. Table II shows the effect of SDS loss in the 8-shot FS-XHumans experiment, which also supports the conclusion of the main text.

Side-view results of the 4-shot setting. In Fig. 4 of the main text, we use a side-view for 2/8-shot tasks to highlight the details of hair reconstruction and another view for the 4-shot task to unveil the SDS effect for unseen regions. To fully present these experiments, we supplement 4-shot side-view results in Fig. VI for comparison.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
$\lambda_{sds} = 0$	26.45	0.9604	0.0343
$\lambda_{sds} = 0.01$	**26.82**	**0.9674**	**0.0301**
$\lambda_{sds} = 1$	25.40	0.9570	0.0375

Table II. The SDS effects on the 8-shot FS-XHumans experiment.

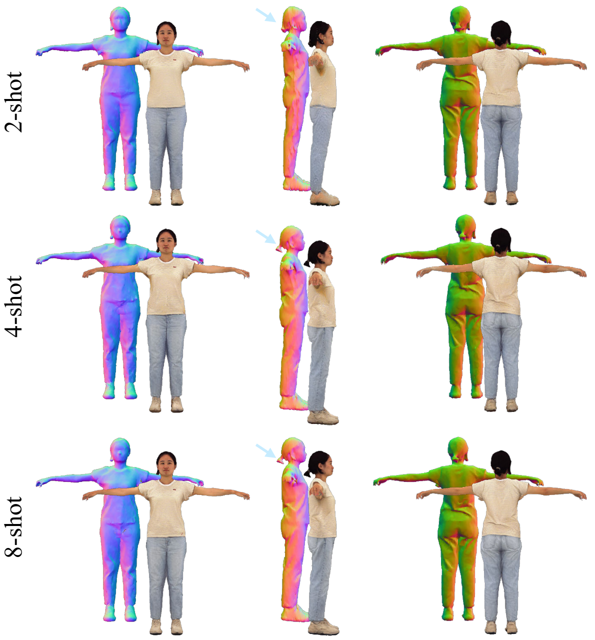

Figure VI. Comparison of body reconstruction with few-shot data

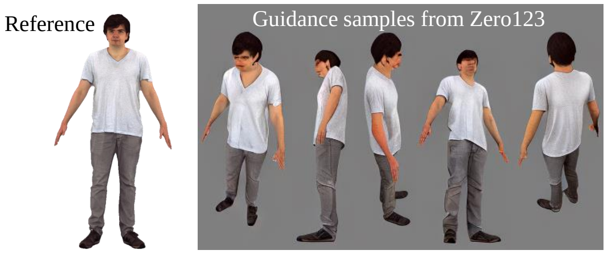

Figure VII. Visualization of Zero123 guidance

The Zero123 guidance. As shown in the Fig. VII, the purely 2D method Zero123 produces low-quality guidance images (*e.g.*, face). Our model achieves performance beyond Zero123 because of a 3D-aware representation and depth/normal supervision.

More results in dynamic demonstration. Please refer to the project page https://seanchenxy.github.io/HaveFunWeb for dynamic results.

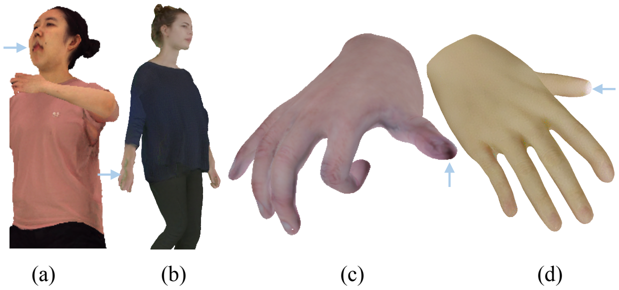

Figure VIII. Demonstration of limitations.

V. Limitations and Future Works

Expression control. To handle varying expressions in training data, we transform expression blendshapes defined by SMPLX [5] into our framework. Nevertheless, the blendshapes are not accurate enough, harming the precision of expression control. The impact on portrait reconstruction is explained in Fig. 7 of the main text. To tackle this difficulty, we will introduce advanced expression control methods (*e.g.*, [7]) to the HaveFun framework.

Disentanglement of albedo and illumination. Our framework generates human texture with mixed albedo and illumination, leading to errors in texture reconstruction. As shown in Fig. VIII(c), some black patterns appear on the top of fingers, which is caused by shadows in the training data. That is, due to a lack of awareness of lighting, the SDS guidance tends to generate shadow-like patterns in unseen regions. To address this issue, we plan to introduce illumination-aware designs (*e.g.*, [2]) to the HaveFun framework.

Full body integration with part-wise few-shot data. This paper streamlines the data collection process and proves that few-shot unconstrained images are cheaper data sources for human avatar creation. In addition, we demonstrate that such a cheap data source is effective for the human body and hand. Nevertheless, we have not used the HaveFun framework for expressive portrait reconstruction. On one hand, because of the aforementioned limitations on facial expression, the HaveFun framework has difficulty in precise expression modeling. In addition, enhancing the accuracy of expression control is far from sufficient for modeling the portrait. For example, because of the lack of inner-mouth regions in the few-shot training data, the avatar is unable to perform a behavior with an open mouth (see Fig. VIII(a)). Therefore, we will explore a few-shot unconstrained data setup for portrait reconstruction. Finally, the portrait, body, and hand can be reconstructed from part-wise few-shot data and integrated into a full representation for an expressive human avatar.

Errors caused by data pre-processing. As illustrated in Fig. VIII, inaccurate image matting results in the introduction of background color to the human texture (Fig. VIII(b)). Additionally, artifacts such as the top of thumb could come from an inaccurate MANO/SMPLX fitting (Fig. VIII(d)).

References

[1] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. 1, 2

[2] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *CVPR*, 2023. 4

[3] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated hand model with diverse accessories and rich textures. In *NeurIPS*, 2022. 1

[4] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Self-Recon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 2

[5] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2, 4

[6] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-Avatar: Expressive human avatars. In *CVPR*, 2023. 1

[7] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*, 2023. 4