

# HiLo: Detailed and Robust 3D Clothed Human Reconstruction with High-and Low-Frequency Information of Parametric Models

## Supplementary Material

### Contents

<b>1. Introduction</b>	<b>1</b>
<b>2. Preliminaries</b>	<b>3</b>
2.1. Signed Distance Function . . . . .	3
2.2. Voxel Grid and Mesh Voxelization . . . . .	3
<b>3. Clothed Human Reconstruction with High- and Low-frequency Information</b>	<b>3</b>
3.1. Progressive Growing High-Frequency SDF . . . . .	3
3.2. Low-Frequency Information for Robustness . . . . .	4
<b>4. Experiments</b>	<b>5</b>
4.1. Toy Experiments . . . . .	6
4.2. Comparison Experiments . . . . .	7
4.3. Ablation Studies . . . . .	8
4.4. Further Discussions . . . . .	8
<b>5. Conclusion</b>	<b>8</b>
<b>6. Acknowledgement</b>	<b>8</b>
<b>A Related Works</b>	<b>1</b>
<b>B More Details of our HiLo</b>	<b>2</b>
B.1. Novelty and differences from previous methods [18, 45, 53, 64] . . . . .	2
B.2. Details on Spatial Interaction MLP . . . . .	2
B.3. Future work and limitations. . . . .	2
<b>C More Experimental Details</b>	<b>2</b>
C.1. Implementation Details . . . . .	2
C.2. More details on Metrics . . . . .	2
C.3. 3D Points Sampling . . . . .	2
C.4. Details of Variant Methods . . . . .	3
C.5. Variant Methods . . . . .	4
<b>D More Details on Datasets</b>	<b>4</b>
<b>E More Experiments</b>	<b>4</b>
E.1. More results on SMPL-X noise. . . . .	4
E.2. More error measurements to assess robustness. . . . .	4
E.3. Is HiLo efficient and light-weighted? . . . . .	4
<b>F More visualization Results</b>	<b>5</b>
F.1. Transfer Sketch to 3D model . . . . .	5
F.2. Results on In-the-wild Images . . . . .	5

### A. Related Works

**Explicit-shape-based approaches** rely on parametric human body models, *e.g.*, SCAPE [4], SMPL [28], SMPL-X [41] to reconstruct 3D humans. Many works [1, 2, 7, 30, 52, 65] introduce the concept of "body+offset", where clothing geometry is represented as 3D displacements on top of the SMPL models. For example, MGN [7] proposes a top-down objective function to align the segmentation maps

of predicted garments and SMPL. To improve the expression ability of garment templates and support more topologies, BCNet [23] disentangles the skinning weight of the garment from the body mesh. Different from the representation of "body+offset", alternative parametric methods adapt vertex deformations on body mesh to capture cloth details. For example, HMD [65] presents the hierarchical deformation framework to recover a detailed human body shape from an initial SMPL mesh in a coarse-to-fine manner. The advantage of these methods lies in their compatibility with the current animation pipeline and ease of control through pose parameters. However, they have limitations in modeling various and complex clothing topologies due to the inherent topology constraints imposed by parametric models.

**Implicit-function-based approaches** aims to reconstruct detailed surfaces with arbitrary topology [12, 32, 37]. This is achieved through the implicit functions, which can be used to approximate 3D representation such as occupancy fields or signed distance fields. PIFu [44] is the pioneering method that utilizes pixel-aligned features for the regression of the occupancy field of human shape. PIFuHD [45] incorporates a multi-level architecture and additional normals to improve the geometric details of PIFu. However, these two methods lack constraints on the global topology of humans, leading to performance degradation in challenging poses. Many works attempt to address this issue in different ways, such as introducing a coarse shape of volumetric humans [18], leveraging depth information of RGB-D images [14]. Unlike the above methods, alternative implicit-function-based methods learn the latent representation of clothing to control the generation of clothing [13, 27, 34]. For example, SMPLicit [13] reconstructs the clothed human by optimizing the latent space of the clothing model to control clothing cut and style. However, the reconstructed human still does not align well with the input image and lacks geometric details.

**Explicit shape & Implicit function approaches** leverage human body models and implicit functions to harness the benefits of both worlds [8, 9, 21]. For instance, PaMIR [64] regularizes the free-form implicit function by incorporating semantic features from the SMPL model. ICON[53], on the other hand, regresses shapes from locally queried features to generalize to unseen poses in in-the-wild photos. ECON [54] combines estimated 2.5D front and back surfaces with an underlying 3D parametric body for improved reconstruction. To further address the variations

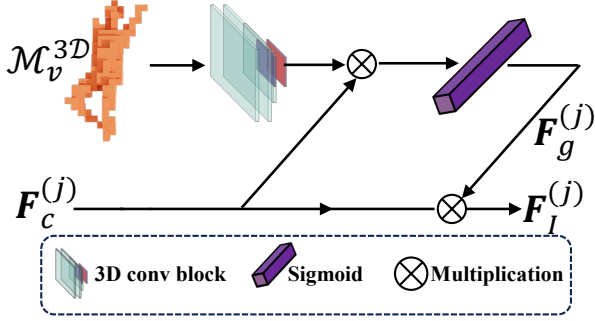


Figure A. Illustration of the spatial interaction module  $\mathcal{A}$ .

in distribution among different spatial points, D-IF [55] introduces a distribution to express the uncertainty of clothing. However, these approaches may fall short when performing highly detailed and robust reconstruction. Specifically, PaMIR is sensitive to global pose and lacks robustness to unseen poses [53]. ECON is prone to reconstruct combined or broken limbs due to the need to complete different surfaces. ICON and D-IF tend to fail in reconstructing detailed parts such as elbows and wrinkles in clothes (see Fig. 1). To promote a detailed reconstruction, our HiLo uses a progressive high-frequency function to improve the expression of reconstructed details. At the same time, HiLo uses the low-frequency-based spatial interactive implicit function to enhance robustness to unseen shapes and poses.

## B. More Details of our HiLo

### B.1. Novelty and differences from previous methods [18, 45, 53, 64]

Clothed human reconstruction from a single RGB image is challenging due to limited views and the absence of depth information. Most recent methods [53–55, 64] rely on parametric body models estimated from RGB images, but they may incur an oversmooth problem due to the **underutilization** of high-frequency (HF) details. Moreover, these methods can be **sensitive** to noises incurred by parametric body model estimation for challenging poses.

To address the above issues, we first **enhance** HF information from the body models to describe geometry details. To this end, we design a progressive growing function to achieve accurate reconstruction while alleviating the convergence difficulty associated with HF information. Moreover, we verify low-frequency (LF) information from the parametric model is **insensitive** to noise. Considering this, we establish a spatial interaction function to leverage the (LF) for robustness reconstruction.

### B.2. Details on Spatial Interaction MLP

Our spatial interaction implicit function takes  $\mathbf{F}_c^1$  that contains our high-frequency SDF  $\mathcal{H}(s; \beta)$ , low-frequency voxel

grids feature  $\mathcal{M}_v^{3D}(\mathbf{p})$ , and normal features  $\mathbf{F}_n(\mathbf{p})$  as input and infers occupancy fields  $\hat{\mathcal{O}}$ .

$$\mathbf{F}_c^1 = [\mathcal{H}(s; \beta), \mathcal{M}_v^{3D}(\mathbf{p}), \mathbf{F}_n(\mathbf{p})] \quad (\text{A})$$

$$\phi_{si}(\mathbf{F}_c^1) \rightarrow \hat{\mathcal{O}}, \quad \phi_{si}(\cdot) = \mathcal{A}^{N+1} \circ T^{(N+1)} \circ \dots \circ \mathcal{A}^1(\cdot) \circ T^{(1)}. \quad (\text{B})$$

As shown in Fig. A, take the 1-th layer of  $\phi_{si}$  as an example, we use attention module [20, 25]  $\mathcal{A}^1$  that takes in the  $\mathcal{M}_v^{3D}$  and  $\mathbf{F}_c^{(1)}$  and output a spatial interaction feature map  $\mathbf{F}_I^{(1)}$ . Specifically, We first extract a global spatial features  $\mathbf{F}_g^{(j)}$  of the  $\mathcal{M}_v^{3D}$  via a 3D Convolution block and a sigmoid function. We achieve the spatial interaction process of different voxels through the equation  $\mathbf{F}_c^{(j)} \times \mathbf{F}_g^{(j)} \rightarrow \mathbf{F}_I^{(j)}$ . After obtaining the  $\mathbf{F}_I^{(1)}$ , we fed it to the first full-connected layer  $T^{(1)}$  to obtain  $\mathbf{F}_c^2$ .

### B.3. Future work and limitations.

**Q5. Future work and limitations.** Since HiLo is trained on orthographic views, it struggles with strong perspectives, causing asymmetrical limbs or unrealistic shapes. This issue is worth studying in the future.

## C. More Experimental Details

We demonstrate the inference details of our HiLo in Alg. 1. The 3D point set is obtained via a coarse-to-fine manner as illustrated in Sec. C.3.

### C.1. Implementation Details

Especially, the dimension of HFSDf, batch size  $b$ , sampled points number  $n$ , and variable dimension channels  $C$  of the spatial interaction module are set to 10, 2, 8000, [39, 512, 256, 128, 1] respectively. The training and testing phases are performed on a single NVIDIA GeForce RTX 3090 GPU. See more details on the training and inference of HiLo in the appendix.

### C.2. More details on Metrics

Specifically, **P2S** denotes the distance between randomly sampled points from a ground truth mesh to its nearest surface on a reconstructed mesh. **Chamfer** is regarded as a bidirectional P2S distance, which computes the distance between randomly sampled points from the reconstructed mesh to its nearest surface on the ground truth mesh. **Normals** is calculated by measuring L2 error between normal images rendered from reconstructed and ground-truth meshes from fixed viewpoints.

### C.3. 3D Points Sampling

During training, we randomly query 3D points inside, outside, and around the SMPL-X surface. During inference,

---

**Algorithm 1:** The inference pipeline of HiLo.

---

- Input:** Sampled 3D points  $\{\mathbf{p}\}_{i=1}^n$ , an RGB image  $\mathcal{I}$  of human, a spatial interactive implicit function  $\phi_{si}$ , a parametric body model estimation net  $E_p$ , a progressive high frequency function  $\mathcal{H}(\cdot; \beta)$ , a 3D CNN  $f_{3D}$ , a mesh voxelization operation  $\mathcal{V}$ , a marching cubes operation  $\mathcal{MC}$ .
- Output:** Triangular mesh of the human.
- 1 Obtaining parametric body model SMPL-X  $\mathcal{M}$  with  $E_p(\mathcal{I})$ .
  - 2 With  $\mathcal{M}$ , obtaining the global voxel grid  $\mathcal{M}_v^{3D}$  using  $f_{3D}(\mathcal{V}(\mathcal{M}))$ .
  - 3 **for**  $\mathbf{p}_i$  **in**  $\{\mathbf{p}\}_{i=1}^n$  **do**
  - 4     Generating SDF  $s$  w.r.t.  $\mathbf{p}_i$  using Eqn. 1.
  - 5     Using  $\mathcal{H}(\cdot; \beta)$  to enhance the SDF  $s$  resulting in point-wise progressive high-frequency SDF  $\mathcal{H}(s; \beta)$ .
  - 6     Obtaining the local voxel grid of  $\mathcal{V}$  by indexing  $\mathcal{M}_v^{3D}$  with  $\mathbf{p}_i$ , resulting in  $\mathcal{V}(\mathbf{p}_i)$ .
  - 7     Get 3D normal features  $\mathbf{F}_n(\mathbf{p}_i)$  w.r.t.  $\mathbf{p}_i$  following ICON.
  - 8     Concatenate  $\mathcal{H}(s; \beta)$ ,  $\mathcal{V}(\mathbf{p}_i)$ ,  $\mathbf{F}_n(\mathbf{p}_i)$ , getting  $F_c^1$ .
  - 9     Using  $\phi_{si}$  to obtaining occupancy field  $\hat{\mathcal{O}}(\mathbf{p}_i)$  from  $F_c^1$  and  $\mathcal{M}_v^{3D}$ , following Eqn. (7).
  - 10 **end**
  - 11 Obtaining the triangular mesh of the human using marching cubes algorithm with  $\mathcal{MC}(\hat{\mathcal{O}})$ .
- 

we define the coordinates of 3D points through an initial 3D grid, and iteratively interpolate the 3D grid to sample 3D points in a more detailed scale.

## C.4. Details of Variant Methods

### C.4.1 Revisit Existing Methods

**PIFu.** To reconstruct a 3D-clothed human, PIFu proposes Pixel-Aligned Implicit Functions to predict whether each 3D point is inside or outside a human surface. Specifically, PIFu learns a 2D feature map from a single image  $I$  using a 2D image encoder via  $f_{2D}(\mathcal{I}) \rightarrow \mathcal{F}_I^{2D}$ . To query local pixel-aligned features on  $\mathcal{F}_I^{2D}$ , PIFu projects 3D points  $\mathbf{p}$  to a 2D plane with  $\pi$  operation and uses bilinear interpolation operation  $S$  to sample the local features from  $\mathcal{F}_I^{2D}$ . The local feature  $f_{2D}(\mathcal{I})(\mathbf{p})$  and the Z coordinate of  $\mathbf{p}$  (*i.e.*,  $\mathbf{p}_z$ ) are concatenated and fed to a multi-layer perceptron (MLP) to obtain the final prediction  $\hat{\mathcal{O}}$ . The pipeline of PIFU follows an equation:

$$\text{PIFu} : \phi(f_{2D}(\mathcal{I})(\mathbf{p})), \mathbf{p}_z \rightarrow \hat{\mathcal{O}}(\mathbf{p}) \quad (\text{C})$$

where  $f_{2D}$  denotes the 2D image encoder. Although PIFu is able to reconstruct high-quality human mesh for commonly seen poses such as walking and standing, PIFu often fails when encountering severe occlusions and large pose variations due to insufficient information from a single image

only.

**PaMIR.** To further regulate the reconstruction process, PaMIR introduces the strengths of parametric body models by learning a parametric-aligned 3D feature volume acquired from a parametric body model, *i.e.*, SMPL. Specifically, PaMIR estimates a SMPL model  $\mathcal{M}$  from the given single image  $I$ , converting  $\mathcal{M}$  to occupancy volume with mesh voxelization  $\mathcal{V}$  and encoding the volume with 3D convolutional neural networks  $f_{3D}$ . Given the voxel-aligned volume features  $f_{3D}(\mathcal{V}(\mathcal{M}), \mathbf{p})$  and the corresponding pixel-aligned feature vector  $f_{2D}(\mathcal{I})(\mathbf{p})$  of  $\mathbf{p}$ , PaMIR learns an implicit function to predict whether  $\mathbf{p}$  is inside or outside a human surface. The pipeline of PaMIR follows the equation:

$$\text{PaMIR} : \phi((f_{2D}(\mathcal{I})(\mathbf{p})), \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p}) \quad (\text{D})$$

Although PaMIR typically feeds their implicit-function module with features of a global 2D image or 3D voxel encoder, but these features are sensitive to global pose [53].

**ICON.** To improve the robustness to out-of-distribution poses, ICON replaces the global encoder of existing methods with a more local scheme: using signed distance function (SDF), barycentric surface normal and local normal features of SMPL regarding  $\mathbf{p}$ . The pipeline of ICON follows the equation:

$$\text{ICON} : \phi(s(\mathbf{p}), \mathcal{F}_n) \rightarrow \hat{\mathcal{O}}(\mathbf{p}) \quad (\text{E})$$

where  $\mathcal{F}_s(\mathbf{p})$  is the signed distance from a query point  $\mathbf{p}$  to the closest body point  $\mathbf{P}^b \in \mathcal{M}$ , and  $\mathcal{F}_n^b$  is the barycentric surface normal of  $\mathbf{P}^b$ , and  $\mathcal{F}_n^c$  is a normal vector. We denote the concatenation of  $\mathcal{F}_n^b(\mathbf{p})$ ,  $\mathcal{F}_n^c(\mathbf{p})$  as  $\mathcal{F}_n$ .

**D-IF.** To alleviate the uncertainty in the process of reconstructing a clothed human, D-IF follows ICON to estimate the occupancy field of the clothed human based on the equation:

$$\begin{aligned} \text{D-IF} : \hat{\mathcal{O}}_f &= \hat{\mathcal{O}}_c + \phi_r(\hat{\mathcal{O}}_c \oplus \mathcal{F}_{7D} \oplus P_\varphi(\mathcal{F}_{7D})) \\ \mathcal{F}_{7D} &= s \oplus \mathcal{F}_n, \hat{\mathcal{O}}_c = \phi(\mathcal{F}_{7D}) \\ \hat{\mathcal{O}}_c(p) &\sim P_\varphi(F_{7D}(\mathbf{p})) = \mathcal{N}(\mu_\varphi(\mathbf{p}), \sigma_\varphi(\mathbf{p})) \end{aligned} \quad (\text{F})$$

where  $\oplus$  denotes concatenate operation,  $P_\varphi(F_{7D}(\mathbf{p}))$  is a Gaussian distribution.

## C.5. Variant Methods

Based on the grasp of existing methods, we introduce the variant methods in our experiments.

$$\text{ICON}_{\text{w}} \mathcal{M}_v^{3D}(p) : \phi(s(\mathbf{p}), \mathcal{F}_n, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})$$

$$\text{D-IF}_{\text{w}} \mathcal{M}_v^{3D}(p) : \phi(\mathcal{F}_{7D}, \mathcal{M}_v^{3D}(\mathbf{p})) + \phi_r(\hat{\mathcal{O}}_c \oplus \mathcal{F}_{7D} \oplus P_\varphi(\mathcal{F}_{7D}))$$

$$\text{HiLo}_{\text{w/o}} \mathcal{M}_v^{3D}(p) : \phi_{si}(\mathcal{H}(s; \beta), \mathcal{F}_n) \rightarrow \hat{\mathcal{O}}(\mathbf{p})$$

$$\text{HiLo}_{\text{w/o}} \mathcal{H}_s(p; \beta) : \phi_{si}(s(\mathbf{p}), \mathcal{F}_n, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})$$

$$\text{HiLo}_{\text{w}} \mathcal{H}_s(p) : \phi_{si}(\mathcal{H}(s), \mathcal{F}_n, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})$$

$$\text{HiLo}_{\text{w/o}} \phi_{si} : \phi(\mathcal{H}(s; \beta), \mathcal{F}_n, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})$$

$$\text{HiLo}_{\text{w/o}} \mathcal{H}(s; \beta) \text{ w/o } \phi_{si} : \phi(s(\mathbf{p}), \mathcal{F}_n, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p}) \quad (\text{G})$$

## D. More Details on Datasets

Our data-split configuration aligns with the protocols outlined by ICON and D-IF. We conduct experiments on the basis of two distinct settings.

- Setting 1: Train on Thuman2.0, test on CAPE. For this setting, we employ 500 scans from Thuman2.0 for training, accompanied by 5 scans for validation. To assess reconstruction accuracy on CAPE, we utilize 150 scans, further categorized into challenging poses (“CAPE-NFP” - 100 scans) and fashion poses (“CAPE-FP” - 50 scans). To emulate diverse viewpoints during testing, RGB images are synthesized by rotating a virtual camera around the textured scans at angles of  $0^\circ$ ,  $120^\circ$ ,  $240^\circ$ .
- Setting 2: Train and test on the same dataset. In this scenario, when training and testing on Thuman2.0, we employ 500 scans for training and reserve 20 scans for testing. Conversely, when training and testing on CAPE, we utilize 120 scans for training, 5 for validation, and 25 for testing.

## E. More Experiments

### E.1. More results on SMPL-X noise.

**SMPL-X Model.** Skinned Multi-Person Linear-Expressive model (SMPL-X) [41] represents human body shapes and poses in a compact and parametric manner. The core idea behind SMPL is to use a linear combination of body shape parameters and joint rotations to represent a 3D human body model with  $N=10475$  vertices and  $K=54$  joints. Specifically, SMPL-X is defined by  $\mathcal{M}(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$ , where  $\theta \in \mathbb{R}^{3(K+1)}$  represents the pose parameter,  $\beta \in \mathbb{R}^{|\theta|}$  is the shape parameter, and  $\psi$  denotes the facial expression parameters, and  $K$  denotes the number of body joints in addition to a joint for global rotation. By adjusting  $\theta$ ,  $\beta$ ,  $\psi$ , SMPL-X is able to represent a wide variety of human body shapes and

poses. See [41] for more details.

**Adding noise to SMPL-X Model.** We further evaluate the robustness ability of our HiLo against various levels of noise in the shape parameters  $\theta_s$  and pose parameters  $\theta_p$  in parametric models. Our experimental setting follows ICON, which samples a scalar value  $\mu \sim \mathcal{N}(0, 1)$ , scaling the noise with two predefined parameters  $s_1$ ,  $s_2$  to represent various levels of noise. The above procedure follows the equation:

$$\begin{aligned} \theta_s + &= s_1 * \mu \\ \theta_p + &= s_2 * \mu \end{aligned} \quad (\text{H})$$

We set  $\{s_1, s_2\}$  to  $\{0.1, 0.1\}$ ,  $\{0.2, 0.2\}$ ,  $\{0.3, 0.3\}$ ,  $\{0.4, 0.4\}$ ,  $\{0.5, 0.5\}$  for a thorough study on the robustness of our HiLo and our baselines. Since we have provided the results w.r.t.  $\{s_1, s_2\} \in [\{0.1, 0.1\}, \{0.2, 0.2\}, \{0.5, 0.5\}]$  in the main draft, we report the remaining results in Tab. A

### E.2. More error measurements to assess robustness.

To further evaluate the robustness of our HiLo, we calculate *Chamfer*, *P2S* and *Normals* between SMPL-X and reconstructed body models. From Tab. B, our HiLo shows better robustness than existing methods.

Table B. Robustness on CAPE.

Methods	Chamfer ( $\downarrow$ )	P2S ( $\downarrow$ )	Normals ( $\downarrow$ )
PIFu	4.0550	3.3971	0.1915
PIFuHD	6.1345	5.2692	0.2017
PaMIR	0.9800	1.0132	0.0714
ICON	0.8198	0.7799	0.0617
D-IF	0.9111	0.8751	0.0666
ECON	0.9083	0.8701	0.0723
HiLo (Ours)	<b>0.6784</b>	<b>0.6580</b>	<b>0.0480</b>

### E.3. Is HiLo efficient and light-weighted?

**Comparison of inference and training time.** In Tab. C, we compare the inference efficiency by the average inference time to reconstruct 200 single-view images. The inference procedures of PaMIR, ICON, D-IF, and HiLo consist of SMPL-X fitting and cloth refinement. Differently, PIFu’s inference procedure only includes cloth refinement, and ECON includes SMPL-X fitting and Poisson Surface Reconstruction (PSR). In terms of inference efficiency, it is evident that our HiLo demonstrates a competitive performance with PaMIR, ICON and D-IF. However, ECON depends on time-consuming PSR to complete human shape, and all other methods show superior performance to it when inference. We measure training efficiency by the average time spent on 10 epochs on the Thuman2 dataset. D-IF needs to train two MLPs and therefore takes more time. We achieve competitive training efficiency with PIFu, PaMIR and ICON even though we introduce high-frequency and

Methods	$\mathcal{M}_v^{3D}$	SMPL-X Noise=0.3			SMPL-X Noise=0.4		
		CAPE-FP	CAPE-NFP	CAPE	CAPE-FP	CAPE-NFP	CAPE
ICON	✗	4.5134	4.7091	4.7069	5.5864	5.9810	5.9015
ICON w $\mathcal{M}_v^{3D}$	✓	4.2250	4.1215	4.2697	3.3824	3.4722	3.3897
D-IF	✗	3.2462	3.6933	3.5700	3.2462	3.6933	3.5700
D-IF w $\mathcal{M}_v^{3D}$	✓	1.2912	1.8222	1.5995	1.2912	1.8222	1.5995
HiLo w/o $\mathcal{M}_v^{3D}$	✗	3.7060	4.3281	4.1071	4.4435	4.8639	4.7763
HiLo	✓	1.1014	1.5407	1.3552	1.1633	1.7584	1.5132

Table A. Impact of  $\mathcal{M}_v^{3D}$  on different methods in terms of Chamfer Distance. We train the models on Thuman2.0 and test them on CAPE.

low-frequency information simultaneously. ECON lacks this statistic because the authors do not release the training codes. **Comparison of model size.** From Tab. C, with the exception of ECON, the model sizes of existing methods are basically the same. Although ECON is lightweight, it requires time-consuming PSR to complete meshes of human shape.

## F. More visualization Results

### F.1. Transfer Sketch to 3D model

Since our HiLo is robust to in-the-wild images [31, 56], we are able to put it to more applications. We show in Fig. B that our HiLo is able to transfer a sketch image of a clothed human into a 3D model with the help of ControlNet [60]. Specifically, we collect sketch images from Pinterest and use ControlNet to transfer the images to RGB images. The RGB images are then fed to our HiLo to reconstruct 3D model of the corresponding human.

### F.2. Results on In-the-wild Images

We report more comparisons with state-of-the-art methods on in-the-wild images in Fig. C, Fig. D, Fig. E, Fig. F, Fig. G, Fig. H, Fig. I, Fig. J, Fig. K. We render the reconstructed 3D models from four different views, *i.e.*,  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ .

Table C. Comparing training/inference efficiency and model size of existing methods.

Method	Inference Time (seconds)	Training Time (seconds)	Million Parameters (seconds)
PIFu [44]	8.13	1636	28.09
PaMIR [64]	21.97	1298	28.18
ICON [53]	18.63	1697	28.11
D-IF [55]	18.51	2336	28.79
ECON [54]	110.93	-	12.07
HiLo (Ours)	19.17	1918	28.21

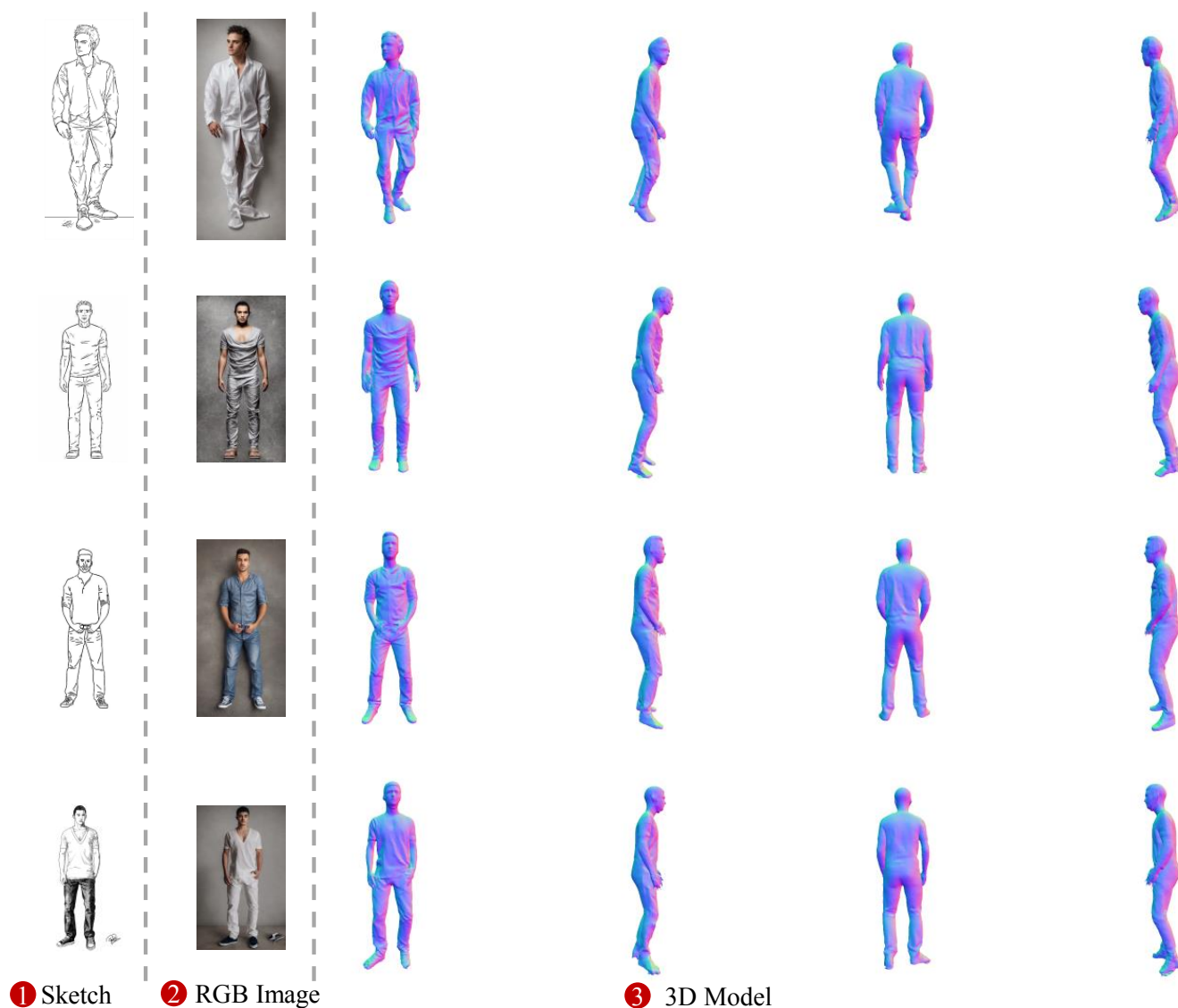


Figure B. More application of our HiLo. We are able to transfer a sketch of a clothed human into a 3D model.

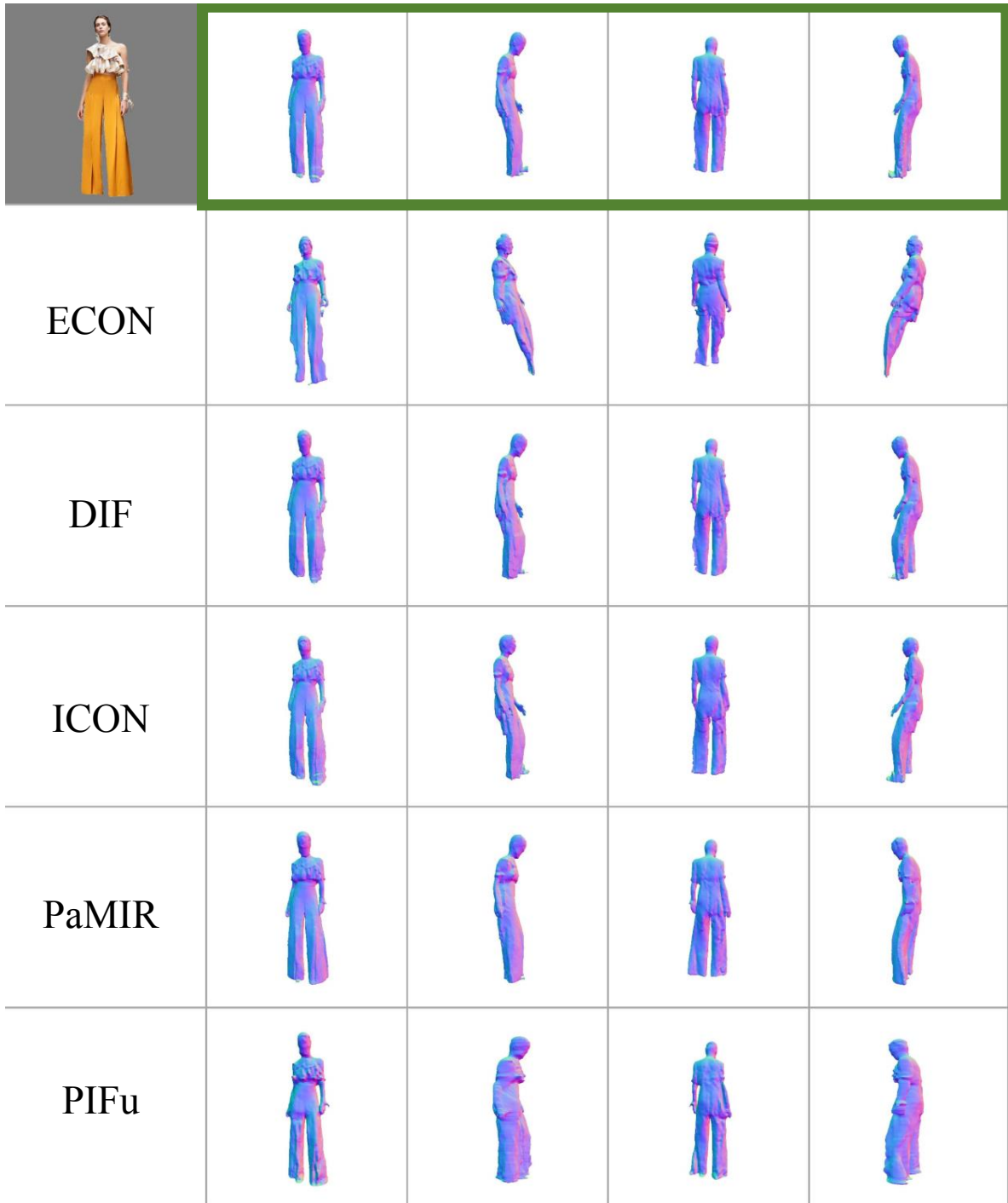


Figure C. Visualization comparisons of reconstruction for our HiLo vs SOTA.

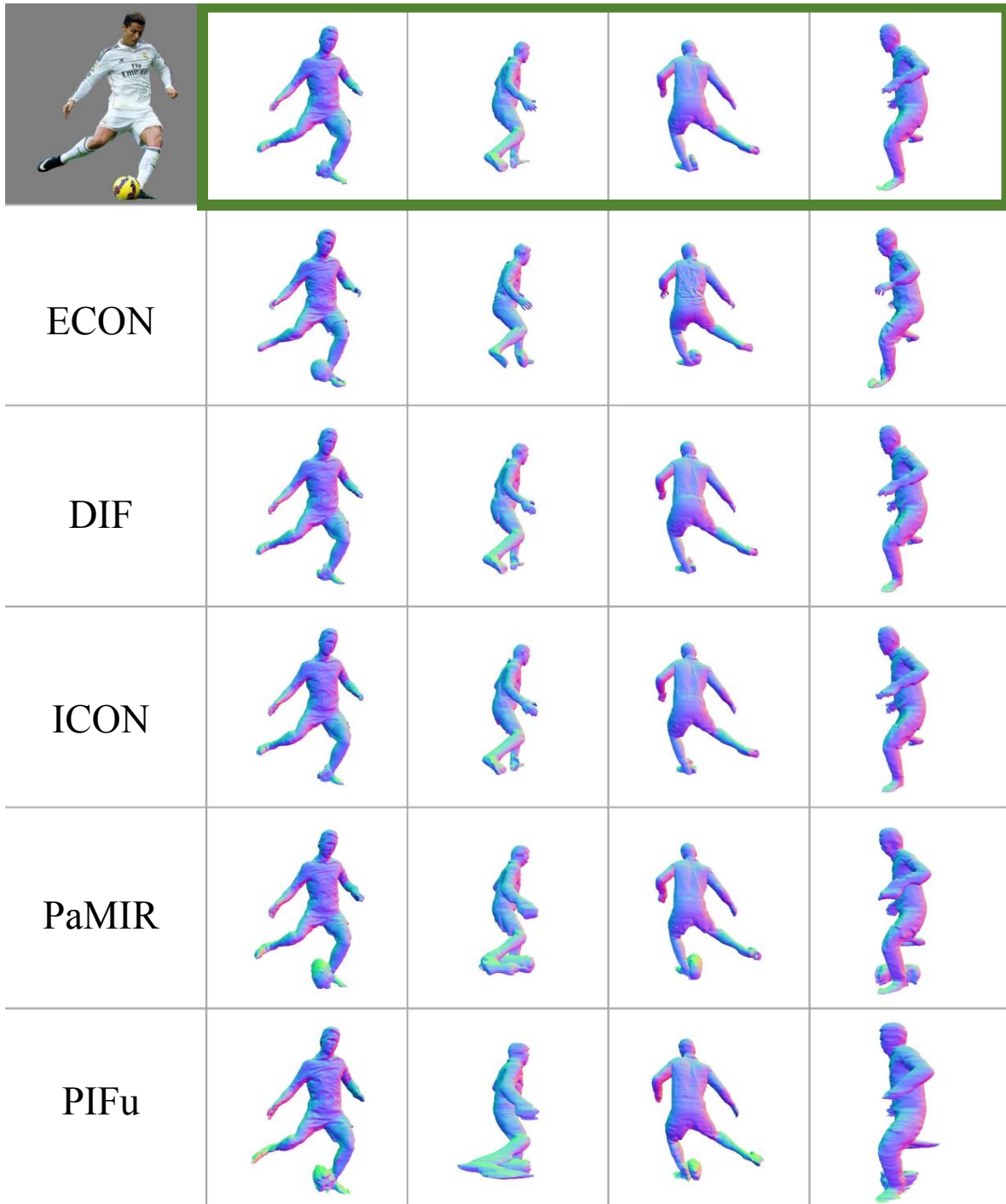


Figure D. Visualization comparisons of reconstruction for our HiLo vs SOTA.



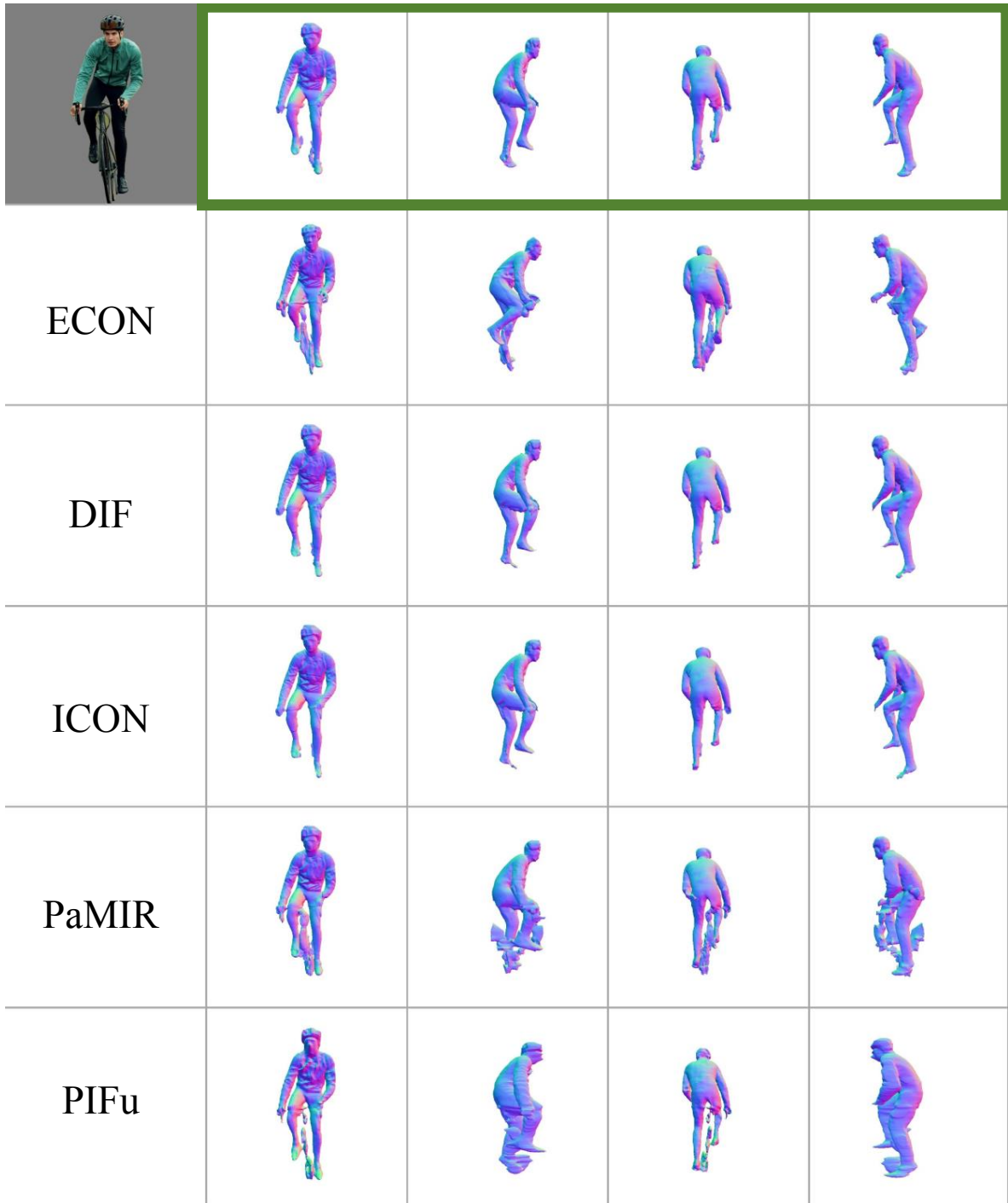


Figure E. Visualization comparisons of reconstruction for our HiLo vs SOTA.

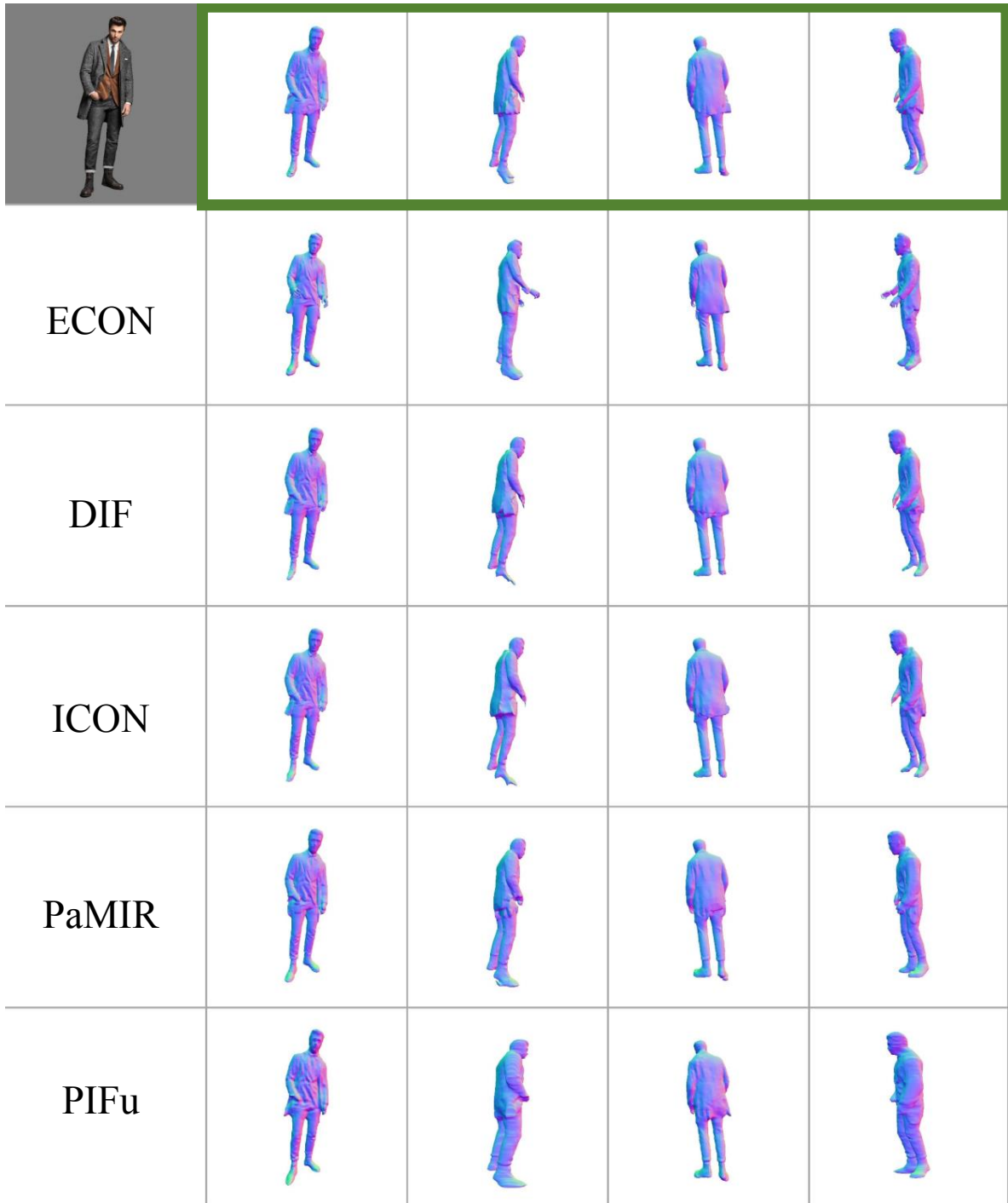


Figure F. Visualization comparisons of reconstruction for our HiLo vs SOTA.





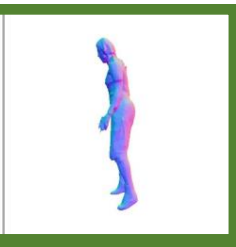

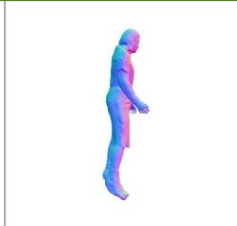

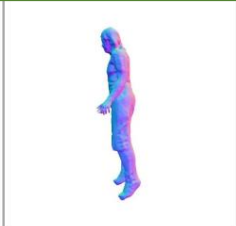




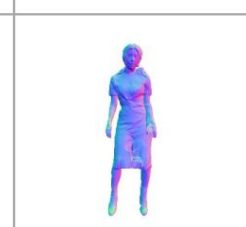


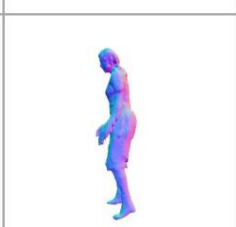

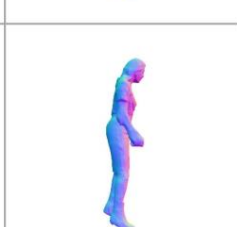

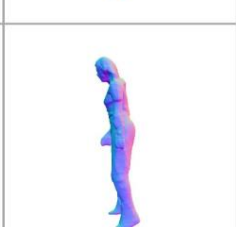

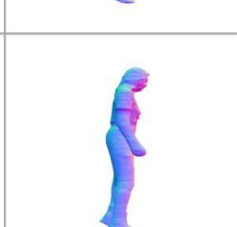

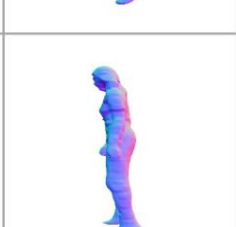
				
ECON				
DIF				
ICON				
PaMIR				
PIFu				

Figure G. Visualization comparisons of reconstruction for our HiLo vs SOTA.
































































































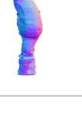




									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				
									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				

Figure H. Visualization comparisons of reconstruction for our HiLo vs SOTA.





































































































									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				
									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				

Figure 1. Visualization comparisons of reconstruction for our HiLo vs SOTA.












































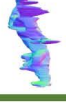
























































									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				
									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				

Figure J. Visualization comparisons of reconstruction for our HiLo vs SOTA.












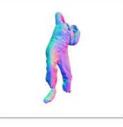











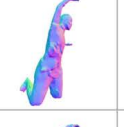












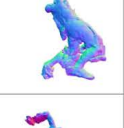


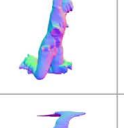
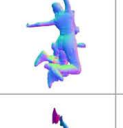



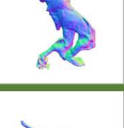


















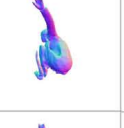


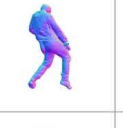




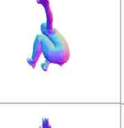
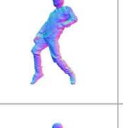

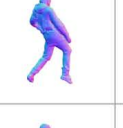


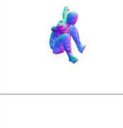
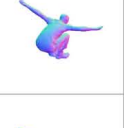
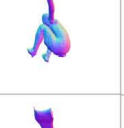
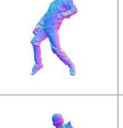




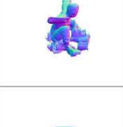



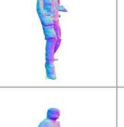
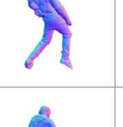



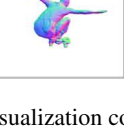
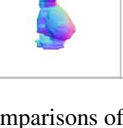
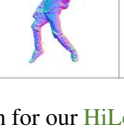



									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				
									
ECON					ECON				
DIF					DIF				
ICON					ICON				
PaMIR					PaMIR				
PIFu					PIFu				

Figure K. Visualization comparisons of reconstruction for our HiLo vs SOTA.