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

Yifan Yang[1,2*] Dong Liu[1*] Shuhai Zhang[1,2] Zeshuai Deng[1] Zixiong Huang[1] Mingkui Tan[1,2,3†]
[1]South China University of Technology [2]Pazhou Lab
[3]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
{seyoungyif, sesmildong, mszhangshuhai, sedengzeshuai, sesmilhzx}@mail.scut.edu.cn
mingkuitan@scut.edu.cn

Abstract

Reconstructing 3D clothed human involves creating a detailed geometry of individuals in clothing, with applications ranging from virtual try-on, movies, to games. To enable practical and widespread applications, recent advances propose to generate a clothed human from an RGB image. However, they struggle to reconstruct detailed and robust avatars simultaneously. We empirically find that the high-frequency (HF) and low-frequency (LF) information from a parametric model has the potential to enhance geometry details and improve robustness to noise, respectively. Based on this, we propose HiLo, namely clothed human reconstruction with high- and low-frequency information, which contains two components. 1) To recover detailed geometry using HF information, we propose a progressive HF Signed Distance Function to enhance the detailed 3D geometry of a clothed human. We analyze that our progressive learning manner alleviates large gradients that hinder model convergence. 2) To achieve robust reconstruction against inaccurate estimation of the parametric model by using LF information, we propose a spatial interaction implicit function. This function effectively exploits the complementary spatial information from a low-resolution voxel grid of the parametric model. Experimental results demonstrate that HiLo outperforms the state-of-the-art methods by 10.43% *and* 9.54% *in terms of Chamfer distance on the Thuman2.0 and CAPE datasets, respectively. Additionally, HiLo demonstrates robustness to noise from the parametric model, challenging poses, and various clothing styles.* [1]

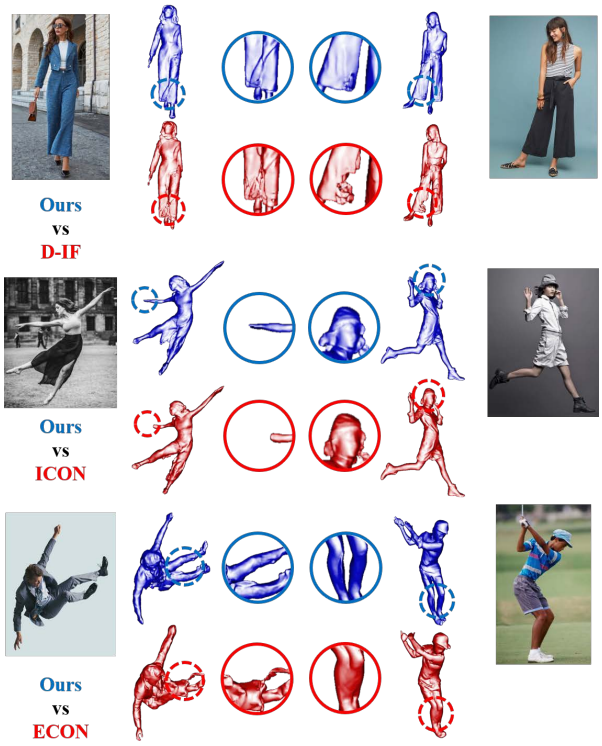

Figure 1. Visualization comparisons on in-the-wild images, our HiLo achieves more accurate and detailed reconstruction on challenging poses and diverse clothes.

1. Introduction

The creation of 3D realistic digital human plays a pivotal role in the realm of mixed reality [23, 33, 53], remote presentation [5, 53], film [16, 46], and gaming [48]. Traditional methods often require expensive and specialized equipment combined with complex artistic efforts to customize the avatars[17, 43, 58], which limits the ability of individuals to create personalized avatars easily. To address the limitation, recent approaches [1, 2, 25, 28, 31, 41, 42, 50–52, 60] capture a 3D avatar from an RGB image of a clothed human, thus eliminating the need for costly scanning equip-

†Corresponding author, *Equal contribution.
[1]Code link: https://github.com/YifYang993/HiLo.git

ment and making it easier for a broader range of users to create personalized avatars.

Despite the convenience of recent advances, the input image usually lacks details about delicate human body parts and diverse clothes from multiple angles. Moreover, the limited viewpoints and lack of accurate depth information make the reconstruction vulnerable to noise, *e.g.,* inaccurate shape and pose of the estimated parametric body model [15, 54, 55]. Therefore, a detailed and robust 3D human reconstruction from an RGB image is challenging.

In spite of the impressive results of the previous methods, they have not fully addressed the problem of *detailed* and *robust* reconstruction simultaneously. Specifically, PIFu [41] produces overly smoothed or non-human body shapes on the unseen side of the human from the input image. ECON [51] requires Poisson Surface Reconstruction and replacement of body parts, introducing an extent of computational overhead (*c.f.* Sec. 4.4). Additionally, there is a risk of body part misalignment when the mid-term data is inaccurate. Considering that clothes need to conform to the surface of naked bodies, the geometry of the parametric model provides effective semantic regularization for reconstructing clothed humans. PaMIR [60], ICON [50], and D-IF [52] use parametric human bodies [26, 38] to regularize the reconstruction. However, the performance of these methods degrades significantly when facing noise on the parameters from the estimated naked bodies (*c.f.* Sec. 4.4).

To achieve *robust* 3D clothed human reconstruction with *detailed* geometry, we aim to explore how to further use the regularization from the parametric model to facilitate this goal. Our exploration is based on two common observations. First, *high-frequency (HF) information* enhances details [30, 40]. Considering that Signed Distance Function (SDF) [34] describes the geometry of a parametric model by representing a distance to the object surface boundary, we investigate the effectiveness of SDF in improving the geometry details of clothed humans. Second, *low-frequency (LF) information* is relatively robust to noise [10, 11, 24, 57]. Since inaccurate parametric model estimation within an error range has an insignificant impact on the corresponding low-resolution voxel grid [45], we seek to use the voxel grid to mitigate the noise of the estimated body. As shown in Fig. 2, qualitative results demonstrate that SDF boosts the reconstruction details while the voxel grid improves robustness to noise. However, how to effectively combine high- and low-frequency information to generate details and mitigate noise simultaneously is still an open question.

In this paper, we propose a high- and low-frequency paradigm *HiLo*, which stands for *clothed human reconstruction with high- and low-frequency information*. To achieve HF detail, we further enhance SDF with HF function [40]. Intuitively, by amplifying the variation of adjacent points that share similar SDFs, we allow for better

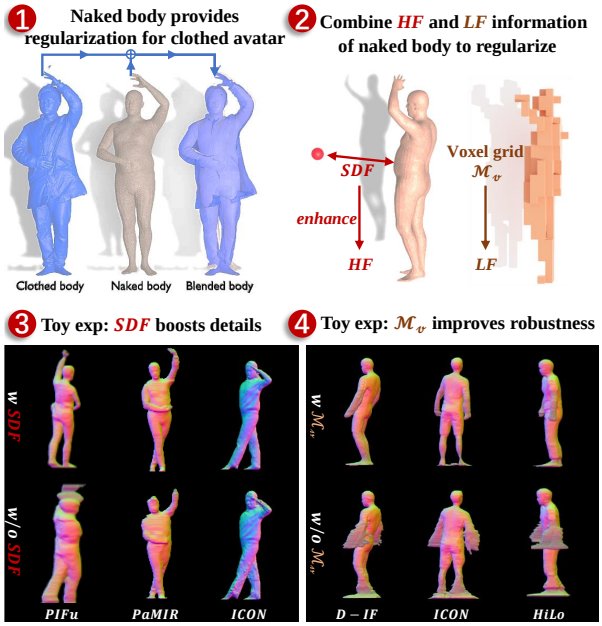

Figure 2. We empirically demonstrate the effectiveness of the high-frequency (HF) regularization from naked bodies in enhancing geometry details in Toy Experiment ❸. We also verify the effectiveness of the low-frequency (LF) regularization in improving robustness to noise in Toy Experiment ❹.

delineation and capturing of fine details in the 3D human. Moreover, to alleviate the convergence difficulty caused by the large gradients amplified by HF function (*c.f.* Sec. 3), we introduce a progressive HF SDF that learns detailed 3D geometry in a coarse-to-fine manner. To achieve robustness, we seek to capture the LF complementary information of the low-resolution voxel grid from the naked human body. To this end, we design a spatial interaction implicit function, which promotes the interaction of global and local information across different voxels via an attention mechanism.

We qualitatively and quantitatively evaluate our HiLo on in-the-wild images and benchmark datasets. The experimental results verify the superiority of HiLo over previous approaches in three key aspects: 1) 3D geometry details (see Fig. 1). 2) Robust reconstruction. 3) Convergence speed. We summarize our contributions in three folds:

- To enhance the geometry details and improve robustness against noise during the clothed human reconstruction process, we propose to explore the high-frequency (HF) information and low-frequency (LF) information from a parametric body model simultaneously.
- To facilitate detailed reconstruction, we introduce a progressive HF function to enhance the signed distance function (SDF) of a parametric model, providing regularization during the reconstruction process. This function learns an HF SDF in a progressive manner to alleviate the convergence difficulty associated with HF informa-

tion. Experimental results show that HiLo reconstructs a more detailed clothed human.

- To ensure robust reconstruction, we employ LF information of low-resolution voxel grids from the parametric model to regularize the reconstruction. We propose a spatial interaction implicit function that reasons complementary information between different voxels. Experimental results show that HiLo is robust to various levels of noise.

2. Preliminaries

2.1. Signed Distance Function

Signed distance function (SDF) [34] is a continuous function that takes a given spatial point p with spatial coordinate $x \in \mathbb{R}^n$ and outputs the distance $s \in \mathbb{R}$ of the point to the closest point on the surface $\partial\Omega$ of an object Ω:

$$\mathcal{F}_s(\mathbf{p}) = s, \quad s = \begin{cases} d(x, \partial\Omega) & \text{if } x \notin \Omega, \\ -d(x, \partial\Omega) & \text{if } x \in \Omega, \end{cases} \tag{1}$$

where $d(x, \partial\Omega) = \inf_{y \in \partial\Omega}(||x - y||_2)$. The sign of the distance implies whether the point is inside (negative) or outside (positive) of the surface $\partial\Omega$. $s = 0$ denotes the point $\mathbf{p}$ locates on the $\partial\Omega$.

2.2. Voxel Grid and Mesh Voxelization

Voxel Grids $\Omega_v \in \mathbb{R}^{d \times h \times w}$ is a representative data structure for describing a 3D object Ω. Specifically, Ω_v is a three-dimensional matrix of 3D space with depth d, height h and width w. Ω_v is composed of equally distributed and equally sized cube-shaped volumetric elements called voxels. The term voxel is the 3D counterpart to a 2D pixel. The resolution of a voxel grid is determined by the size of the voxels and the dimensions of the grid. Lower resolution implies larger voxels, resulting in a coarser representation of the space. Thus, we use the low-resolution voxel grid to represent the low-frequency information of a 3D object.

Mesh voxelization $\mathcal{V}$ is a computational technique that plays a crucial role in converting irregular continuous 3D geometric models [4, 26, 38] such as triangular mesh and point clouds into regular and discrete voxel grids. In this process, a 3D mesh Ω_m, which is a collection of connected triangles, is converted into a grid of voxels $\Omega_v \in \mathbb{R}^3$. In this paper, we use $\mathcal{V}$ to transfer SMPL-X mesh $\mathcal{M}$ to a low-resolution voxelized mesh $\mathcal{M}_v \in \mathbb{R}^{32 \times 32 \times 32}$, and then use a 3D CNN to encode $\mathcal{M}_v$ for $\mathcal{M}_v^{3D} \in$ for more flexibility. Based on our observation from Tab. 3, the low-frequency natural of $\mathcal{M}_v^{3D}$ aids multiple existing methods in mitigating various noise levels in SMPL-X shape and pose.

3. Clothed Human Reconstruction with High- and Low-frequency Information

We aim to robustly infer detailed 3D clothed avatars from RGB images $\mathcal{I}$. Recent advances [8, 9, 19, 20] tend to use parametric naked body $\mathcal{M}$ such as SMPL [26] or SMPL-X [38] estimated from $\mathcal{I}$ to provide semantic regularization on clothed human avatars. We empirically verify that high-frequency (HF) and low-frequency (LF) information from $\mathcal{M}$ are able to refine geometry and improve robustness to the reconstruction of the clothed human (Sec. 4.1). Based on this, we propose clothed human reconstruction with high- and low-frequency information, namely HiLo, which balances the HF and LF information to achieve detailed and robust reconstruction simultaneously. As shown in Fig. 3, HiLo contains two key components: (1) To refine the geometry of clothed human with HF information, we propose to use *progressive high-frequency function* to enhance the signed distance function (SDF) of $\mathcal{M}$ (*c.f.* Sec. 2), and alleviate the convergence difficulty introduced by large gradients in a coarse-to-fine enhancement manner. (2) To achieve robust reconstruction using low-frequency information, we explore complementary information from low-resolution voxels $\mathcal{M}_v$ from $\mathcal{M}$ for a more comprehensive understanding of human geometry. To this end, we design a *spatial interactive implicit function* (*c.f.* Sec. 3.2) that leverages spatial information from local and global voxelized SMPL-X to predict the occupancy field $\hat{\mathcal{O}}$. Finally, we use the Marching Cubes ([27]) to obtain the 3D mesh of the clothed avatar from $\hat{\mathcal{O}}$.

The overall optimization of our proposed HiLo minimizes the following objective function:

$$\mathcal{L}_{overall} = \mathcal{L}_a(\hat{\mathcal{O}}, \mathcal{O}), \tag{2}$$

where $\mathcal{L}_a$ is the MSE loss and $\mathcal{O}$ is a GT occupancy field.

3.1. Progressive Growing High-Frequency SDF

Given that SDFs can enhance 3D reconstruction quality as confirmed in Sec. 4.1, we will leverage this for more realistic avatar reconstruction. However, directly fitting input coordinates with SDFs may lead to subpar representation of *high-frequency* variation in geometry (see Sec. 4.3). This aligns with previous work [39] indicating neural networks prioritize learning *low-frequency* signals. We will explore effective strategies to mitigate this.

Conventional high-frequency SDF. To improve the ability to represent complicated 3D shapes robustly, a straightforward way is to apply periodic functions $\mathcal{H}$ such as sine and cosine [44] to extract high-frequency signals on SDF of each sampled point via

$$\begin{aligned} \mathcal{H}(s) &= [s, \mathcal{H}_0(s), \mathcal{H}_1(s), \ldots, \mathcal{H}_k(s), \ldots, \mathcal{H}_L(s)], \\ \mathcal{H}_k(s) &= [\sin(2^k \pi s), \cos(2^k \pi s)], k \in \{0, 1, \ldots, L\}. \end{aligned} \tag{3}$$

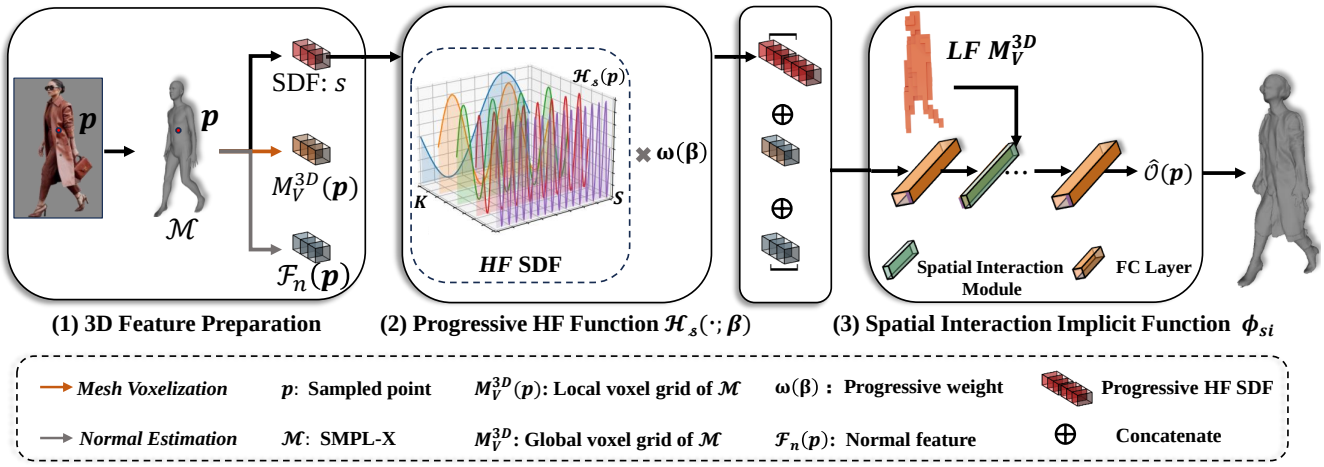

Figure 3. Overview of our proposed HiLo. Conditioned on a single-view image $\mathcal{I}$ and the corresponding SMPL-X $\mathcal{M}$, we first prepare a signed distance field s and a low-resolution voxel grid $\mathcal{M}_v^{3D}$ of the naked body. Then, our proposed progressive high-frequency signed distance function $\mathcal{H}(s;\beta)$ enhances s for detailed geometry of the clothed human and alleviates convergence difficulties introduced by large gradients in a coarse-to-fine learning manner. Moreover, we design an implicit function ϕ_{si} which leverages the complementary information of low-frequency voxels from $\mathcal{M}_v^{3D}$ to mitigate various levels of noise. Finally, we combine the above HF and LF features to ϕ_{si} to infer the occupancy field $\hat{\mathcal{O}}$ of the clothed avatar.

In this way, we amplify the variation of adjacent points that share similar SDFs, allowing for better delineation and capturing of fine details in the 3D object.

Despite the positive characteristic of high-frequency SDF, effective updating for parameters is difficult. Specifically, the gradient of $\mathcal{H}_k(s)$ w.r.t. s is calculated by

$$\frac{\partial \mathcal{H}_k(s)}{\partial s} = 2^k\pi[\cos(2^k\pi s), -\sin(2^k\pi s)]. \quad (4)$$

Eqn. (4) incorporated with the coefficient $2^k\pi$ will significantly amplify the gradient signals regarding SDF, especially for larger k. Large gradients could lead to convergence difficulties and numerical instability, ultimately resulting in poor representation performance [3, 36].

High-frequency SDF in a growing manner. To address the above issue, we introduce a progressively growing approach as shown in Fig. 3 (2), initially emphasizing low-frequency signal learning and gradually focusing on learning the high-frequency geometry. Specifically, in the early stage of training, we reduce the weight of high-frequency signals (*e.g.*,$\mathcal{H}_k(s)$) which have higher k and progressively increase their importance during training. We formulate this schedule as $\mathcal{H}_k(s;\beta)$ with a weight $\omega_k(\beta)$ following [35]:

$$\mathcal{H}_k(s;\beta) = \omega_k(\beta)[\sin(2^k\pi s), \cos(2^k\pi s)],$$
$$\omega_k(\beta) = \begin{cases} 0 & \text{if } \beta - k < 0; \\ \frac{1-\cos((\beta-k)\pi)}{2} & \text{if } 0 \le \beta - k < 1; \\ 1 & \text{if } \beta - k > 1, \end{cases} \quad (5)$$

where β is proportional to the iteration of the optimization process, see Fig. 4 for the relationship between $\omega_k(\beta)$ and β. With $\omega_k(\beta)$, the gradient of $\mathcal{H}_k(s;\beta)$ becomes

$$\frac{\partial \mathcal{H}_k(s;\beta)}{\partial s} = \omega_k(\beta)2^k\pi[\cos(2^k\pi s), -\sin(2^k\pi s)]. \quad (6)$$

Then, during the beginning of the optimization, β is set so small that only frequency components with a smaller value of k will be assigned a non-zero weight, while the frequency components with a higher value of k will be omitted. Throughout the optimization, the higher-frequency components are progressively activated. This manner allows HiLo to explore the low-frequency part and later focus on the fine-grained geometry of 3D humans.

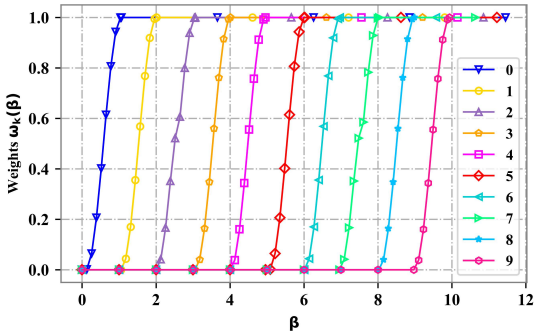

Figure 4. Illustration of the relationship between progressive weights ω and β during the training process.

3.2. Low-Frequency Information for Robustness

Most recent methods [50–52, 60] are based on SMPL-X. However, SMPL-X estimation often faces **misalignment is-**

sues with the corresponding image, especially when facing a challenging human pose. Thus, it is crucial to **achieve robust reconstruction** against misaligned SMPL-X. Our results (*c.f.* Fig. 2) show that the low-frequency information, represented by low-resolution voxel grids $\mathcal{M}_v$ of SMPL-X $\mathcal{M}$, enhances robustness against noise. We leverage this insight to incorporate local and global information of $\mathcal{M}_v$ for improved regularization in reconstruction.

Local voxels for 3D feature preparation. Motivated by that point-wise local 3D features from $\mathcal{M}_v$ are robust to out-of-distribution pose and shape [60], we voxelize the estimated SMPL-X $\mathcal{M}$ and query it by $\mathbf{p}$. Specifically, to obtain the voxelization features, we convert the corresponding SMPL-X $\mathcal{M}$ into a low-resolution voxel grid $\mathcal{M}_\mathcal{V}$ by mesh voxelization operation $\mathcal{V}$ [60] and encode it via a 3D CNN f_{3D} for a 3D feature volume $\mathcal{M}_\mathcal{V}^{3D}$. To obtain point-wise 3D features, we use trilinear interpolation to sample $\mathcal{M}_v^{3D}$ based on coordinate $\mathbf{p}$ of sampled 3D points, resulting in $\mathcal{M}_\mathcal{V}^{3D}(\mathbf{p})$. We empirically find that by combining $\mathcal{M}_\mathcal{V}^{3D}(\mathbf{p})$, HiLo is robust to SMPL-X noise (*c.f.* Sec. 4.4). As shown in Fig. 3 (1), in addition to $\mathcal{H}(s;\beta)$ and $\mathcal{M}_\mathcal{V}^{3D}$, we follow ICON [50] to use a normal feature $\mathbf{F}_\mathrm{n}(\mathbf{p})$ to provide detailed texture information. Then, we concatenate them into one final feature $\mathbf{F}_c^1$=$[\mathcal{H}(s;\beta), \mathcal{M}_\mathcal{V}^{3D}(\mathbf{p}), \mathbf{F}_\mathrm{n}(\mathbf{p})]$ and then fed $\mathbf{F}_c^1$ to our designed implicit function to reconstruct the clothed avatar.

Global voxels for spatial interaction implicit function. To reconstruct clothed avatars, a typical solution is to map 3D features $\mathbf{F}$ to a continuous occupancy field that represents the interior and exterior of a clothed human. To this end, numerous literature [41, 50, 52, 60] uses an implicit function parameterized by a memory-efficient multi-layer perceptron (MLP) $\mathcal{T}$ to map $\mathbf{F}$ into an occupancy field $\hat{\mathcal{O}}$.

However, the potential issue of the existing implicit function lies in its underutilization of the global information inherent in 3D data. Previous research [6, 21, 32] has shown that the representation ability of features can be improved by capturing the global correlation between the features. For 3D clothed human reconstruction, different human body parts contain distinct yet complementary spatial information. For instance, as shown in Fig. 5 (a), the voxels located on the shoulder (the red point) may offer valuable topological cues to constrain the prediction of geometry near the elbow (the blue point).

To leverage global information from the voxel grid of SMPL-X of $\mathcal{M}_v^{3D}$, we design a spatial interaction module $\mathcal{A}$ into ϕ to infer the 3D occupancy, denoted by ϕ_{si}, see detail in Appx B.2. As shown in Fig. 5 (b), ϕ_{si} injects global-scale features of $\mathcal{M}_v^{3D}$ to the local 3D feature with the aim of introducing whole-body awareness to ϕ_{si}:

$$\phi_{si}(\mathbf{F}_c^1) \rightarrow \hat{\mathcal{O}},\ \ \phi_{si}(\cdot) = \mathcal{A}^{N+1} \circ T^{(N+1)} \circ \cdots \circ \mathcal{A}^1(\cdot) \circ T^{(1)}. \tag{7}$$

Optimization. We optimize parameters of ϕ_{si} and f_{3D} via minimizing the MSE loss in Eqn. (2) between the predicted occupancy field $\hat{\mathcal{O}}$ and the ground-truth occupancy field $\mathcal{O}$. With $\hat{\mathcal{O}}$, we reconstruct the triangular mesh of the 3D clothed avatar via marching cubes algorithm [27].

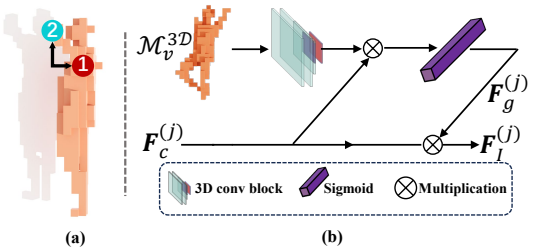

Figure 5. (a) Complementarity of voxel ❶ and ❷. (b) Illustration of the spatial interaction module $\mathcal{A}$.

4. Experiments

Datasets: We conduct experiments on two open-source datasets, *i.e.,* Thuman2.0 [59] and CAPE [28], which both contain various human shapes with different human poses and diverse clothes. Specifically, following ICON [50], the CAPE dataset is divided into the "CAPE-FP" and "CAPE-NFP" sets, which have "fashion" and "non-fashion" poses, respectively, to further analyze the generalization to complex body poses. Moreover, to evaluate our HiLo on in-the-wild images, we follow ICON to collect 200 diverse images from Pinterest[2]. The images contain humans performing dramatic movements or wearing diverse clothes.

Metrics: We evaluate our HiLo and baseline methods in terms of three metrics: **Chamfer distance** and **P2S distance** mainly measure coarse geometry error, while **Normals** mainly captures high-frequency differences. See details in Appx. C.2.

Baselines: We compare our proposed HiLo with mainstream state-of-the-art methods, including PIFu [41], PaMIR [60], ICON [50], ECON [51] and D-IF [52], refer to the Appx. C.4 for the detailed description. To demonstrate the necessity of naked 3D body regularization, we first conduct a toy experiment to study the effect of SDF on different baselines. To this end, we design three variants based on PIFu, PaMIR, and ICON, which incorporate SDF into existing methods (PIFu and PaMIR) or remove SDF from the existing method (ICON), namely $\text{PIFu}_{\text{w SDF}}$, $\text{PaMIR}_{\text{w SDF}}$ and $\text{ICON}_{w/o\ \text{SDF}}$, respectively.

For ablation studies, we construct several variants of our HiLo: 1) $\text{HiLo}_{w/o\ \phi_{si}}$ replaces our spatial interaction implicit function with the vanilla implicit function; 2) $\text{HiLo}_{w/o\ \mathcal{M}_v^{3D}}$ is constructed by removing the voxelized SMPL-X; 3) $\text{HiLo}_{w/o\ \mathcal{H}(s;\beta)}$ is constructed by replacing our progressive HF SDF with vanilla SDF.

[2]https://www.pinterest.com/

Figure 6. Visualization results of 3D clothed avatar reconstruction with our HiLo from in-the-wild images, which present various clothing and challenging poses. We show the front (blue) and rotated (red) views.

Table 1. Toy experiments about the impact of SDF on 3D clothed human reconstruction, on seen (*i.e.,* training and test on the same dataset) and unseen (*i.e.,* training on Thuman2.0 and test on CAPE) settings.

Dataset	CAPE-FP			CAPE-NFP			CAPE			Thuman2			CAPE		
Methods	Chamfer (↓)	P2S (↓)	Normals (↓)	Chamfer	P2S	Normals	Chamfer	P2S	Normals	Chamfer	P2S	Normals	Chamfer	P2S	Normals
PIFu [41]	2.1000	2.0930	0.0910	2.9730	2.9400	0.1110	2.6820	2.6580	0.1040	2.6880	2.5730	0.0970	7.1244	2.7633	0.3902
$\text{PIFu}_{\text{w SDF}}$	0.8908	0.8637	0.0676	0.9848	0.9545	0.0698	0.9437	0.9178	0.0707	1.6659	1.7934	0.1360	1.3078	1.4306	0.0980
PaMIR [60]	1.2250	1.2060	0.0550	1.4130	1.3210	0.0630	1.3500	1.2830	0.0600	1.4388	1.5613	0.1071	0.9339	0.9444	0.0659
$\text{PaMIR}_{\text{w SDF}}$	0.9188	0.8788	0.0565	1.1132	1.0729	0.0611	1.0112	0.9725	0.0601	1.4073	1.5624	0.1174	0.8438	0.8179	0.0572
ICON [50]	0.7475	0.7488	0.0508	0.8656	0.8639	0.0545	0.8055	0.8084	0.0539	1.1431	1.3020	0.0923	0.8610	0.8878	0.0606
$\text{ICON}_{\text{w/o SDF}}$	1.0243	0.9478	0.0741	1.4862	1.3313	0.0919	1.2736	1.1538	0.0850	1.3114	1.2116	0.1015	7.4892	1.7708	0.3780

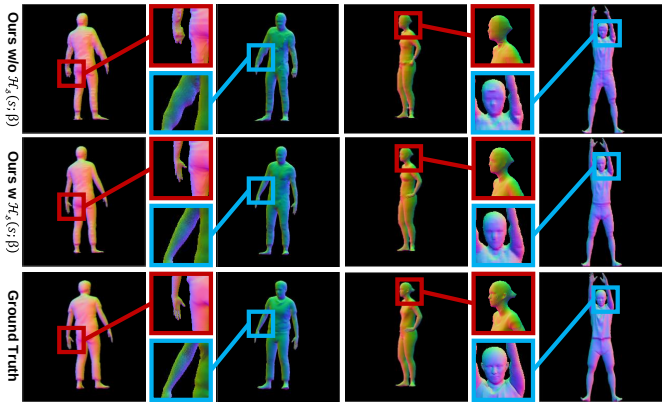

Figure 7. Reconstruction results with or without our progressive high-frequency SDF $\mathcal{H}(s;\beta)$. The geometry details of clothes, hands, faces are enhanced by introducing $\mathcal{H}(s;\beta)$.

Implementation details: We implement our approach using PyTorch[3] [37] and train our networks with RMSprop [47] optimizer. For a fair comparison, we follow all common hyper-parameter settings same as ICON [50]. See more implementation details in the Appx.

[3]We will release our code.

4.1. Toy Experiments

Our motivation stems from the idea that HF information and LF information improve details and robustness, respectively. To verify this, we employ two tools, *i.e., SDF* and *voxelized SMPL-X* $\mathcal{M}_v^{3D}$ to establish this constraint.

Impact of SDF. We build upon three representative methods, *i.e.,* PIFu, PaMIR, and ICON for the experiments. Specifically, $\text{PIFu}_{w\ \text{SDF}}$ adds SDF following equations: $\phi(\mathcal{F}_s(\mathbf{p}), f_{2\text{D}}(\mathcal{I})(\mathbf{p})) \rightarrow \hat{\mathcal{O}}$. $\text{PaMIR}_{w\ \text{SDF}}$ incoorporates SDF following: $\phi(\mathcal{F}_s(\mathbf{p}), f_{2\text{D}}(\mathcal{I})(\mathbf{p}), \mathcal{V}(\mathcal{M})(\mathbf{p})) \rightarrow \hat{\mathcal{O}}$. $\text{ICON}_{w/o\ \text{SDF}}$ replaces SDF with the z coordinate of $\mathbf{p}$ following: $\phi(\mathcal{F}_{\text{n}}^{\text{b}}(\mathbf{p}), \mathcal{F}_{\text{n}}^{\text{c}}(\mathbf{p}), \mathbf{p}_z) \rightarrow \hat{\mathcal{O}}$, where ϕ is the vanilla implicit function and f_{2D} is a 2D CNN. Experimental results in Fig. 2, and Tab. 1 demonstrate the SDF improves geometry details compared with variant methods without it.

Impact of voxelized SMPL-X M_v^{3D}**.** Our empirical verification is based on ICON, D-IF and HiLo that requires SMPL-X for reconstruction. we design three variants named $\text{ICON}_{\text{w}\ \mathcal{M}_v^{3D}(p)}$, $\text{D-IF}_{\text{w}\ \mathcal{M}_v^{3D}(p)}$ and $\text{HiLo}_{\text{w/o}\ \mathcal{M}_v^{3D}(p)}$ that add or remove M_v^{3D}. From the experimental results in Fig. 2 and Tab. 3, we find that incorporating $\mathcal{M}_v^{3D}$ helps achieve a more robust reconstruction even faces various levels of noise in SMPL-X shape and pose. The reason is that the low-resolution voxel grid of the

Table 2. (A) Comparison experiments and (B) ablation studies on seen (*i.e.,* training and test on the same dataset) and unseen (*i.e.,* training on Thuman2.0 and test on CAPE) settings. The **bold** and the underlined numbers indicate the best and second-best results, respectively. "-" denotes that PIFuHD and ECON does not provide a training code.

Group		Train on Thuman2.0 and test on CAPE									Train and test on the same dataset (Thuman2.0 or CAPE)					
	Dataset	CAPE-FP			CAPE-NFP			CAPE			Thuman2.0			CAPE		
	Methods	Chamfer (↓)	P2S (↓)	Normals (↓)	Chamfer	P2S	Normals	Chamfer	P2S	Normals	Chamfer	P2S	Normals	Chamfer	P2S	Normals
A	PIFu [41]	2.1000	2.0930	0.0910	2.9730	2.9400	0.1110	2.6820	2.6580	0.1040	2.6880	2.5730	0.0970	7.1244	2.7633	0.3902
	PIFuHD [42]	2.3020	2.3350	0.0900	3.7040	3.5170	0.1230	3.2370	3.1230	0.1120	2.4613	2.3605	0.0924	-	-	-
	PaMIR [60]	1.2250	1.2060	0.0550	1.4130	1.3210	0.0630	1.3500	1.2830	0.0600	1.4388	1.5613	0.1071	0.9339	0.9444	0.0659
	ICON [50]	0.7475	0.7488	0.0508	0.8656	0.8639	0.0545	0.8055	0.8084	0.0539	1.1431	1.3020	0.0923	0.8610	0.8878	0.0606
	ECON [51]	0.9651	0.9175	**0.0412**	0.9983	0.9694	**0.0415**	0.9872	0.9521	**0.0414**	-	-	-	-	-	-
	D-IF [52]	0.8038	0.7766	0.0546	0.9877	0.9491	0.0611	0.8878	0.8574	0.0589	1.0305	1.0864	0.0830	0.8332	0.8489	0.0597
B	$\text{HiLo}_{\text{w/o } \mathcal{H}(s;\beta)}$	0.7564	0.7449	0.0514	0.8697	0.8658	0.0553	0.8118	0.8045	0.0547	0.9442	1.0323	0.0785	0.7971	0.7999	0.0551
	$\text{HiLo}_{\text{w/o } \phi_{si}}$	0.7996	0.7860	0.0569	0.9112	0.9042	0.0601	0.8555	0.8449	0.0468	0.9886	1.0836	0.0850	0.7999	0.7948	0.0547
	$\text{HiLo}_{\text{w/o } \mathcal{H}(s;\beta) \text{ w/o } \phi_{si}}$	0.8662	0.8970	0.0647	1.0201	1.0525	0.0706	0.9362	0.9720	0.0690	1.1220	1.2544	0.0954	0.8125	0.8224	0.0588
	HiLo (Ours)	**0.6954**	**0.6876**	0.0471	**0.7830**	**0.7876**	0.0499	**0.7430**	**0.7428**	0.0499	**0.9230**	**0.9855**	**0.0732**	**0.7861**	**0.7729**	**0.0544**

Table 3. Toy experiments on 3D reconstruction with different levels of SMPL-X noise in terms of chamfer distance on unseen CAPE dataset. The voxel grid of naked body $\mathcal{M}_v^{3D}(p)$ improves the robustness of reconstruction.

		SMPL-X Noise=0.1			SMPL-X Noise=0.2			SMPL-X Noise=0.5		
Methods	$\mathcal{M}_v^{3D}(p)$	CAPE-FP	CAPE-NFP	CAPE	CAPE-FP	CAPE-NFP	CAPE	CAPE-FP	CAPE-NFP	CAPE
ICON [50]	✗	1.7949	2.0537	1.9284	2.6695	3.0917	2.9365	4.2181	4.8266	4.6716
$\text{ICON}_{\text{w } \mathcal{M}_v^{3D}(p)}$	✓	1.4381	1.5380	1.3411	2.1950	2.3067	2.0723	2.2129	2.3760	2.1188
D-IF [52]	✗	1.6078	1.7881	1.7037	2.6853	3.1002	2.9962	4.3591	4.8006	4.7310
$\text{D-IF}_{\text{w } \mathcal{M}_v^{3D}(p)}$	✓	1.3557	1.7877	1.6064	1.8022	2.4110	2.1959	3.0307	4.4190	4.0807
$\text{HiLo}_{\text{w/o } \mathcal{M}_v^{3D}(p)}$	✗	1.9518	2.1966	2.0877	2.9315	3.4561	3.2661	4.6382	5.0606	4.9844
HiLo	✓	1.0517	1.3210	1.2004	1.0893	1.5876	1.3427	1.0960	1.6156	1.3593

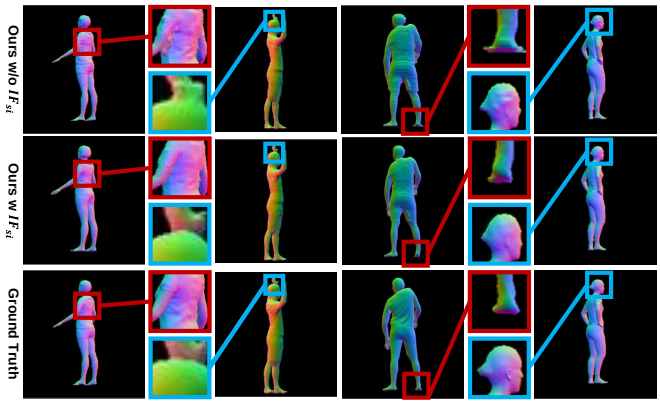

Figure 8. Reconstructions w and w/o our spatial interaction implicit function ϕ_{si}. Our ϕ_{si} is able to perceive the global human body and is therefore able to remove non-human shapes.

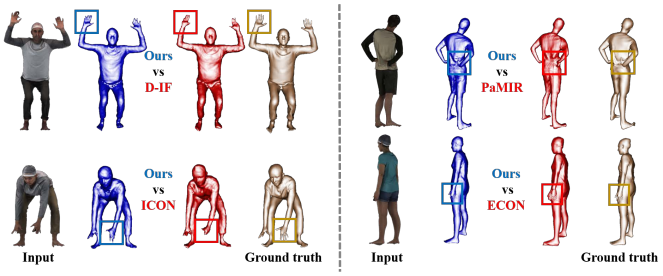

Figure 9. Visualization comparisons on CAPE dataset. The model is training on Thuman2.0 dataset.

naked body is insensitive to noise, and therefore provides robust low-frequency regularization in the training process.

4.2. Comparison Experiments

Quantitative results. We conducted comparative experiments in Tab. 2 under two settings. 1) *Setting1*: Following the setting of the previous methods, we train and test on the same datasets. 2) *Setting2*: To further evaluate the generalization ability of our HiLo on unseen datasets, we train and test HiLo using different datasets. Our approach achieves the best results in the seen and the unseen settings due to the high- and low-frequency paradigm.

Visualization Results. We compare our HiLo with baselines on in-the-wild images and CAPE dataset in Fig. 1, and Fig. 9, respectively. The results show that our HiLo is able to reconstruct 3D clothed avatars with more realistic details. Although ECON obtains detailed fingers by replacing the hand of the SMPL-X model, there exists misalignment on the connection wrist when the corresponding SMPL-X is inaccurate. We put more visualization results of our HiLo on in-the-wild images in Fig. 6. The results demonstrate the ef-

Table 4. Ablations of our progressive high-frequency SDF on 3D reconstruction with different levels of SMPL-X noise in terms of chamfer distance on unseen CAPE dataset. In addition to $\mathcal{M}_v^{3D}$, our progressive high-frequency SDF is also to handle SMPL-X noise due to the coarse-to-fine learning manner.

					SMPL-X Noise=0.1			SMPL-X Noise=0.2			SMPL-X Noise=0.5		
Methods	$\mathcal{M}_v^{3D}(p)$	$\mathcal{H}_s(p;\beta)$	$\mathcal{H}_s(p)$	SDF	CAPE-FP	CAPE-NFP	CAPE	CAPE-FP	CAPE-NFP	CAPE	CAPE-FP	CAPE-NFP	CAPE
HiLo$_{w/o\ \mathcal{H}_s(p;\beta)}$	✓	✗	✗	✓	1.1435	1.4700	1.3124	1.3401	1.9339	1.6909	1.2861	1.8200	1.5620
HiLo$_{w\ \mathcal{H}_s(p)}$	✓	✗	✓	✗	1.1932	1.5541	1.3904	1.1243	1.5575	1.3701	1.2794	1.8973	1.6652
HiLo	✓	✓	✗	✗	1.0517	1.3210	1.2004	1.0893	1.5876	1.3427	1.0960	1.6156	1.3593

fectiveness and generalization ability of our HiLo in recovering detailed geometry (such as hairs and cloth wrinkles). We put more visualization results in the Appx.

4.3. Ablation Studies

How does $\mathcal{H}(s;\beta)$ improve geometry details? We quantitatively demonstrate the necessity of $\mathcal{H}(s;\beta)$ in Tab. 2. The results demonstrate that HiLo$_{w/o\ \mathcal{H}(s;\beta)}$, the variant method that replaces $\mathcal{H}(s;\beta)$ with standard SDF, achieves inferior performance than HiLo. To further study the impact of $\mathcal{H}(s;\beta)$ on common (-FP) and challenging poses (-NFP), we evaluate HiLo on cape dataset that contains both categories. Tab. 2 demonstrates that $\mathcal{H}(s;\beta)$ improves the performance of avatar reconstruction more on challenging poses (9.54% improvement) than in fashion (5.21% improvement in terms of Chamfer distance. Furthermore, Fig. 7 demonstrates that $\mathcal{H}(s;\beta)$ leads to more detailed reconstruction, resulting in clearer arms and more realistic cloth wrinkles. From the results, we observe that incorporating the power of high frequency with SDF helps in capturing detailed geometry.

How does ϕ_{si} improve body topology of the reconstructed avatar? As shown in Fig. 8 and Tab. 2, our ϕ_{si} removes the non-human shape and boosts reconstruction performance. The reason is that our ϕ_{si} leverages a cross-scale attention module $\mathcal{A}$ that builds topological signals between different spatial points in the body model.

4.4. Further Discussions

Is our $\mathcal{H}_s(p;\beta)$ able to help HiLo be robust to SMPL-X noise? We study the impact of $\mathcal{H}_s(p;\beta)$ on the robustness ability of our HiLo by replacing it with conventional high frequency SDF $\mathcal{H}_s(p)$ and vanilla SDF. We perturb the SMPL-X model with various levels of noise to compare the robustness of the SDF variants. As illustrated in Tab. 4, our proposed $\mathcal{H}_s(p;\beta)$ outperforms SDF and high-frequency SDF variants under multiple noise scales due to the progressive manner. See more results in the Appx.

Is HiLo able to converge faster? In the comparison of validation accuracy depicted in Fig. 10, it is evident that our HiLo exhibits a remarkable ability to rapidly converge and attain superior performance. Specifically, HiLo swiftly reaches a commendable accuracy of 0.90 at approximately iteration 100. In contrast, the second best method, *i.e.,* ICON, takes significantly longer, around iteration 200, to reach the same level of accuracy, underscoring the efficiency and efficacy of our approach.

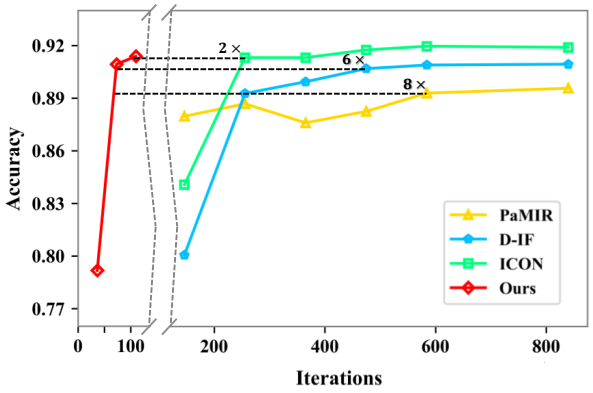

Figure 10. Convergence curves of different methods on CAPE dataset. Our HiLo is able to converge faster than existing methods.

5. Conclusion

In this paper, we propose a high-frequency and low-frequency paradigm by exploiting high-frequency and low-frequency information from parametric body models. Based on the paradigm, we design clothed human reconstruction with high- and low-frequency information, namely HiLo that contains: 1) a progressive high-frequency SDF to improve geometry details and alleviate large gradients that hinder model convergence; 2) a spatial interaction implicit function that utilizes the low-frequency complementary information from the voxelized naked body to improve robustness against noise. Experimental results demonstrate the superiority of our HiLo. In the future, we will apply our method to more 3D reconstruction tasks such as 3D face reconstruction, and indoor scene 3D reconstruction.

6. Acknowledgement

This work was partially supported by National Natural Science Foundation of China (NSFC) 62072190, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183, and TCL Science and Technology Innovation Fund.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 1

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 1

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *Advances in neural information processing systems*, 32, 2019. 4

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 3, 1

[5] Sahar Aseeri and Victoria Interrante. The influence of avatar representation on interpersonal communication in virtual social environments. *IEEE transactions on visualization and computer graphics*, 27(5):2608–2617, 2021. 1

[6] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016. 5

[7] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 1

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *The European Conference on Computer Vision*, pages 311–329, 2020. 3, 1

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 3, 1

[10] Qingwen Bu, Dong Huang, and Heming Cui. Towards building more robust models with frequency bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4402–4411, 2023. 2

[11] Yiting Chen, Qibing Ren, and Junchi Yan. Rethinking and improving robustness of convolutional neural networks: a shapley value-based approach in frequency domain. In *Advances in Neural Information Processing Systems*, pages 324–337, 2022. 2

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 1

[13] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11875–11885, 2021. 1

[14] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20470–20480, 2022. 1

[15] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*, pages 1775–1784, 2018. 2

[16] Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. Continuity editing for 3d animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 1

[17] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics*, 38(6): 1–19, 2019. 1

[18] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *Advances in Neural Information Processing Systems*, 33:9276–9287, 2020. 1, 2

[19] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021. 3

[20] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3, 1

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

[22] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *The European Conference on Computer Vision*, pages 18–35. Springer, 2020. 1

[23] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. Effects of avatar and background representation forms to co-presence in mixed reality (mr) tele-conference systems. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: modelling and simulating virtual humans and environments*, pages 1–4. 2016. 1

[24] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7245–7254, 2020. 2

[25] Ren Li, Benoît Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body. In *Pro-*

ceedings of the Asian Conference on Computer Vision, pages 2780–2795, 2022. 1

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Acm Transactions on Graphics*, 34, 2015. 2, 3, 1

[27] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353, 1998. 3, 5

[28] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 1, 5

[29] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *The European Conference on Computer Vision*, 2020. 2

[31] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *European conference on computer vision*, pages 184–200, 2022. 1

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *The European Conference on Computer Vision*, pages 483–499, 2016. 5

[33] Seung-Tak Noh, Hui-Shyong Yeo, and Woontack Woo. An hmd-based mixed reality system for avatar-mediated remote collaboration with bare-hand interaction. In *Proceedings of the 25th International Conference on Artificial Reality and Telexistence and 20th Eurographics Symposium on Virtual Environments*, pages 61–68, 2015. 1

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2, 3, 1

[35] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 4

[36] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013. 4

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[38] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 3, 1, 4

[39] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 3

[40] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310, 2019. 2

[41] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1, 2, 5, 6, 7

[42] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 7, 2

[43] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *The European Conference on Computer Vision*, 2016. 1

[44] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 3

[45] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2

[46] Lin Sun. Research on the application of 3d animation special effects in animated films: Taking the film avatar as an example. *Scientific Programming*, 2022. 1

[47] Tijmen Tieleman, Geoffrey Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 6

[48] Ea Christina Willumsen. Is my avatar my avatar? character autonomy and automated avatar actions in digital games. In *DiGRA Conference*, 2018. 1

[49] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 1

[50] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13296–13306, 2022. 1, 2, 4, 5, 6, 7, 3

[51] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5, 7, 1, 6

[52] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-if: Uncertainty-aware human digitization via implicit distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9122–9132, 2023. 1, 2, 4, 5, 7, 6

[53] Boram Yoon, Hyung-il Kim, Gun A Lee, Mark Billinghurst, and Woontack Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 547–556. IEEE, 2019. 1

[54] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 2

[55] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 5

[57] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International conference on machine learning*, pages 7502–7511, 2019. 2

[58] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014. 1

[59] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 5

[60] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 1, 2, 4, 5, 6, 7

[61] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4491–4500, 2019. 1

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

Supplementary Material

Contents

A. Related Works

Explicit-shape-based approaches rely on parametric human body models, *e.g.,* SCAPE [4], SMPL [26], SMPL-X [38] to reconstruct 3D humans. Many works [1, 2, 7, 28, 49, 61] 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 [22] 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 [61] 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, 29, 34]. This is achieved through the implicit functions, which can be used to approximate 3D representation such as occupancy fields or signed distance fields. PIFu [41] is the pioneering method that utilizes pixel-aligned features for the regression of the occupancy field of human shape. PIFuHD [42] 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, 25, 31]. 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, 20]. For instance, PaMIR [60] regularizes the free-form implicit function by incorporating semantic features from the SMPL model. ICON[50], on the other hand, regresses shapes from locally queried features to generalize to unseen poses in in-the-wild photos. ECON[51] combines estimated 2.5D front and back surfaces with an underlying 3D parametric body for improved reconstruction. To further address the variations in

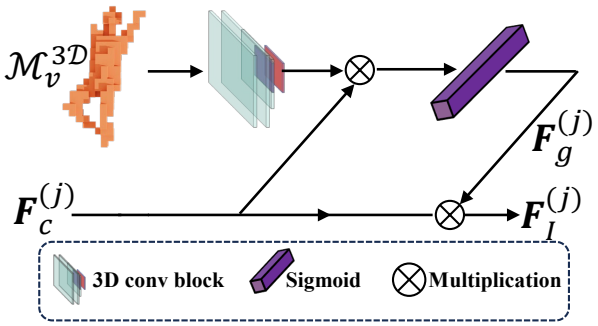

Figure A. Illustration of the spatial interaction module $\mathcal{A}$.

distribution among different spatial points, D-IF[52] 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 [50]. 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, 42, 50, 60]

Clothed human reconstruction from a single RGB image is challenging due to limited views and the absence of depth information. Most recent methods [50–52, 60] 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 $\mathbf{F}_{\mathrm{n}}(\mathbf{p})$ as input and infers occupancy fields $\hat{\mathcal{O}}$.

$$\mathbf{F}_c^1 = [\mathcal{H}(s;\beta), \mathcal{M}_V^{3D}(\mathbf{p}), \mathbf{F}_{\mathrm{n}}(\mathbf{p})] \tag{A}$$

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

As shown in Fig. A, take the 1-th layer of ϕ_{si} as an example, the attention module $\mathcal{A}^1$ 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, an spatial interactive implicit function ϕ_{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.

Obtaining parametric body model SMPL-X $\mathcal{M}$ with $E_p(\mathcal{I})$.
With $\mathcal{M}$, obtaining the global voxel grid $\mathcal{M}_v^{3D}$ using $f_{3D}(\mathcal{V}(\mathcal{M}))$.
for $\mathbf{p}_i$ *in* $\{\mathbf{p}\}_{i=1}^{n}$ **do**
    Generating SDF s w.r.t. $\mathbf{p}_i$ using Eqn. 1.
    Using $\mathcal{H}(\cdot\,;\beta)$ to enhance the SDF s resulting in point-wise progressive high-frequency SDF $\mathcal{H}(s;\beta)$.
    Obtaining the local voxel grid of $\mathcal{V}$ by indexing $\mathcal{M}_v^{3D}$ with $\mathbf{p}_i$, resulting in $\mathcal{V}(\mathbf{p}_i)$.
    Get 3D normal features $\mathbf{F}_{\mathrm{n}}(\mathbf{p}_i)$ w.r.t. $\mathbf{p}_i$ following ICON.
    Concatenate $\mathcal{H}(s;\beta)$, $\mathcal{V}(\mathbf{p})$, $\mathbf{F}_{\mathrm{n}}(\mathbf{p}_i)$, getting F_c^1.
    Using ϕ_{si} to obtaining occupancy field $\hat{\mathcal{O}}(\mathbf{p}_i)$ from F_c^1 and $\mathcal{M}_v^{3D}$, following Eqn. (7).
end
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 $f_{2\mathrm{D}}(\mathcal{I}) \rightarrow \mathcal{F}_I^{2D}$. To query local pixel-aligned features on $\mathcal{F}_I^{2D}$, PIFu projects 3D points $\mathbf{p}$ to a 2D plane with π operation and uses bilinear interpolation operation S to sample the local features from $\mathcal{F}_I^{2D}$. The local feature $f_{2\mathrm{D}}(\mathcal{I})(\mathbf{p})$ 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:

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

where $f_{2\mathrm{D}}$ 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 $f_{3\mathrm{D}}$. Given the voxel-aligned volume features $f_{3\mathrm{D}}(\mathcal{V}(\mathcal{M})), \mathbf{p}))$ and the corresponding pixel-aligned feature vector $f_{2\mathrm{D}}(\mathcal{I})(\mathbf{p}))$ 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:

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

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 [50].

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 PaMIR follows the equation:

$$\mathrm{ICON} : \phi(s(\mathbf{p}), \mathcal{F}_{\mathrm{n}}) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \tag{E}$$

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

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:

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

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

C.5. Variant Methods

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

$$
\begin{aligned}
&\text{ICON}_{\text{w } \mathcal{M}_v^{3D}(p)} : \phi(s(\mathbf{p}), \mathcal{F}_{\text{n}}, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \\
&\text{D-IF}_{\text{w } \mathcal{M}_v^{3D}(p)} : \phi(\mathcal{F}_{\text{7D}}, \mathcal{M}_v^{3D}(\mathbf{p})) + \phi_r(\hat{\mathcal{O}}_c \oplus \mathcal{F}_{\text{7D}} \oplus P_\varphi(\mathcal{F}_{\text{7D}})) \\
&\text{HiLo}_{\text{w/o } \mathcal{M}_v^{3D}(p)} : \phi_{si}(\mathcal{H}(s;\beta), \mathcal{F}_{\text{n}}) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \\
&\text{HiLo}_{\text{w/o } \mathcal{H}_s(p;\beta)} : \phi_{si}(s(\mathbf{p}), \mathcal{F}_{\text{n}}, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \\
&\text{HiLo}_{\text{w } \mathcal{H}_s(p)} : \phi_{si}(\mathcal{H}(s), \mathcal{F}_{\text{n}}, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \\
&\text{HiLo}_{\text{w/o } \phi_{si}} : \phi(\mathcal{H}(s;\beta), \mathcal{F}_{\text{n}}, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p})) \\
&\text{HiLo}_{\text{w/o } \mathcal{H}(s;\beta) \text{ w/o } \phi_{si}} : \phi(s(\mathbf{p}), \mathcal{F}_{\text{n}}, \mathcal{M}_v^{3D}(\mathbf{p})) \rightarrow \hat{\mathcal{O}}(\mathbf{p}))
\end{aligned}
\tag{G}
$$

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 $0°, 120°, 240°$.
- 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) [38] 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 ψ denotes the facial expression parameters, and K denotes the number of body joints in addition to a joint for global rotation. By adjusting θ, β, ψ, SMPL-X is able to represent a wide variety of human body shapes and poses. See [38] 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 θ_s and pose parameters θ_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:

$$
\begin{aligned}
\theta_s &+= s_1 * \mu \\
\theta_p &+= s_2 * \mu
\end{aligned}
\tag{H}
$$

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 (↓)	P2S (↓)	Normals (↓)
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)	**0.6784**	**0.6580**	**0.0480**

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

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.

and ICON even though we introduce high-frequency and low-frequency information simultaneously. ECON lacks this statistic because the authors do not release the training codes. **Comparison of model size.** Form 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, 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 [56]. 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.,* $0°, 90°, 180°, 270°$.

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

Method	Inference Time (seconds)	Training Time (seconds)	Million Parameters (seconds)
PIFu [41]	8.13	1636	28.09
PaMIR [60]	21.97	1298	28.18
ICON [50]	18.63	1697	28.11
D-IF [52]	18.51	2336	28.79
ECON [51]	110.93	-	12.07
HiLo (Ours)	19.17	1918	28.21

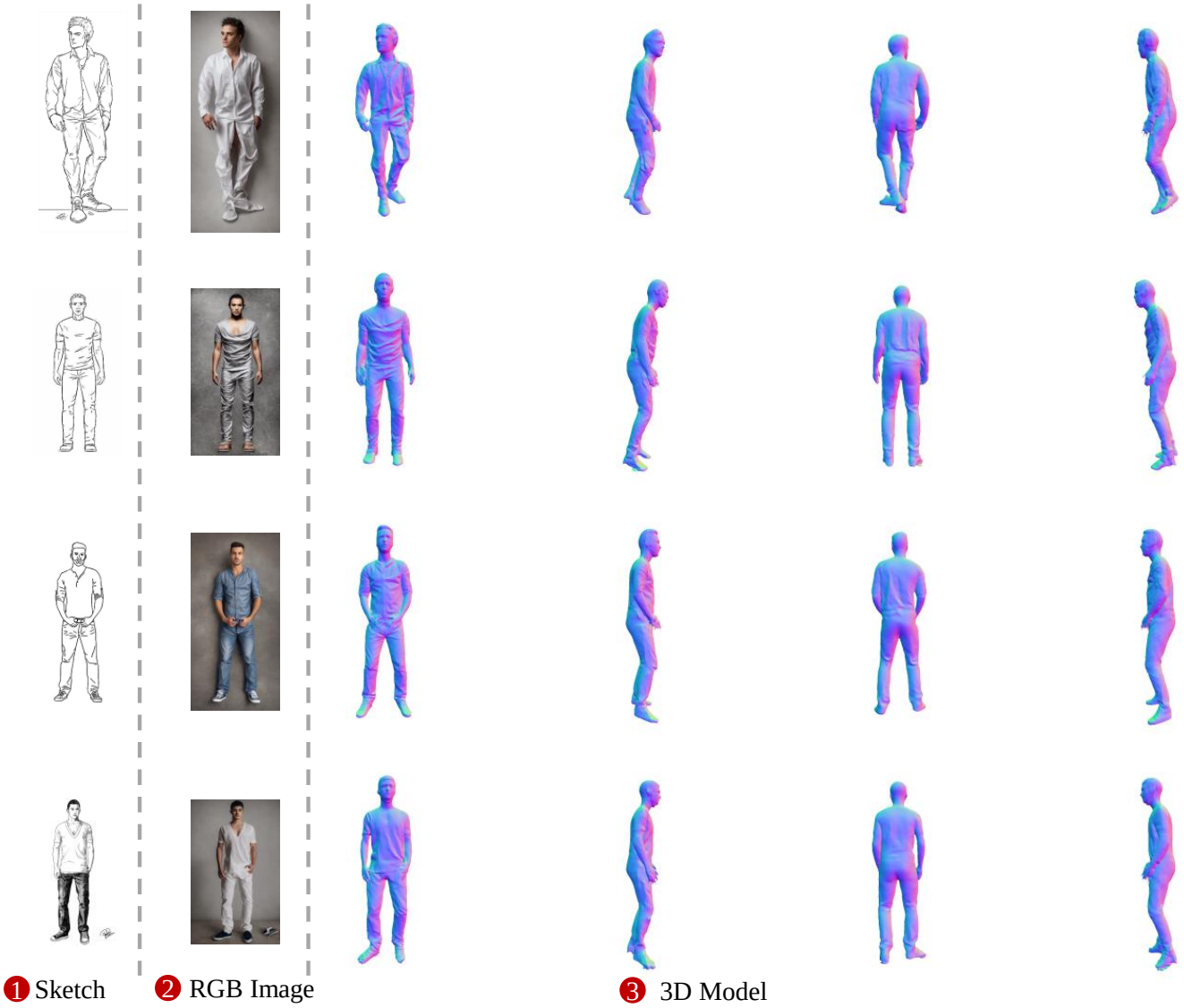

Figure B. More application of our HiLo. We are able to transfer a sketch of a clothed human into a 3D model.

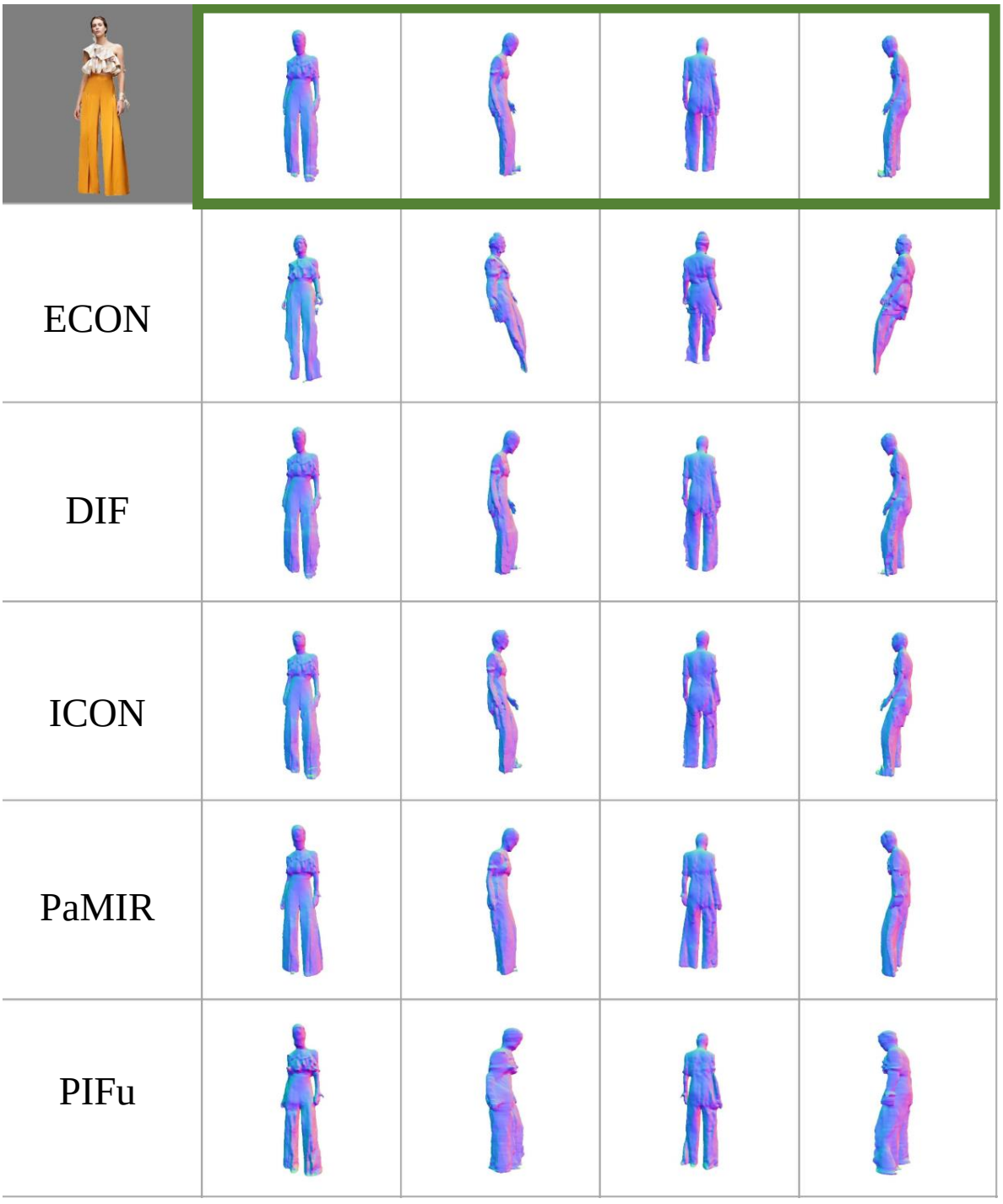

Figure C. Visualization comparisons of reconstruction for our HiLo vs SOTA.

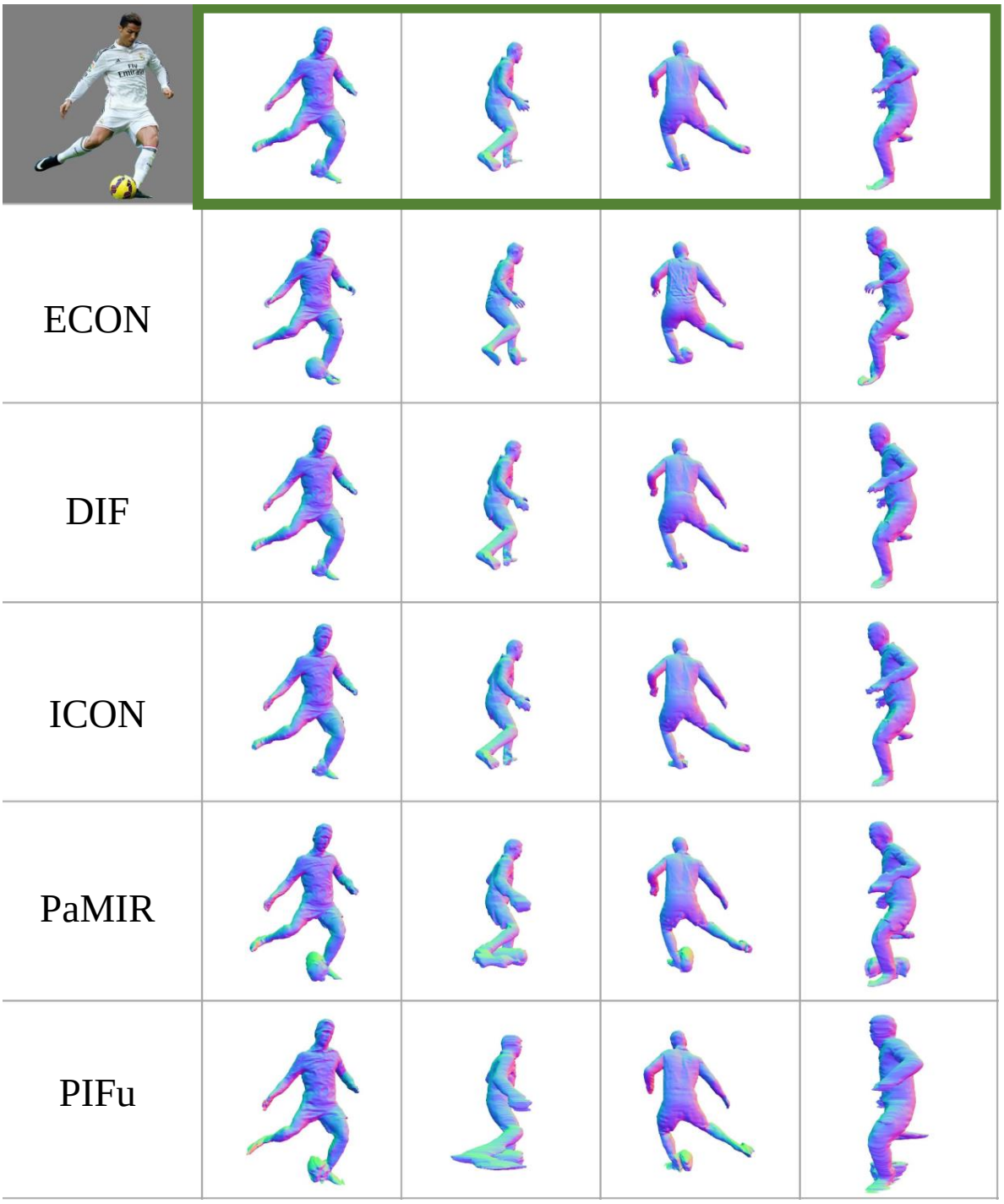

Figure D. Visualization comparisons of reconstruction for our HiLo vs SOTA.

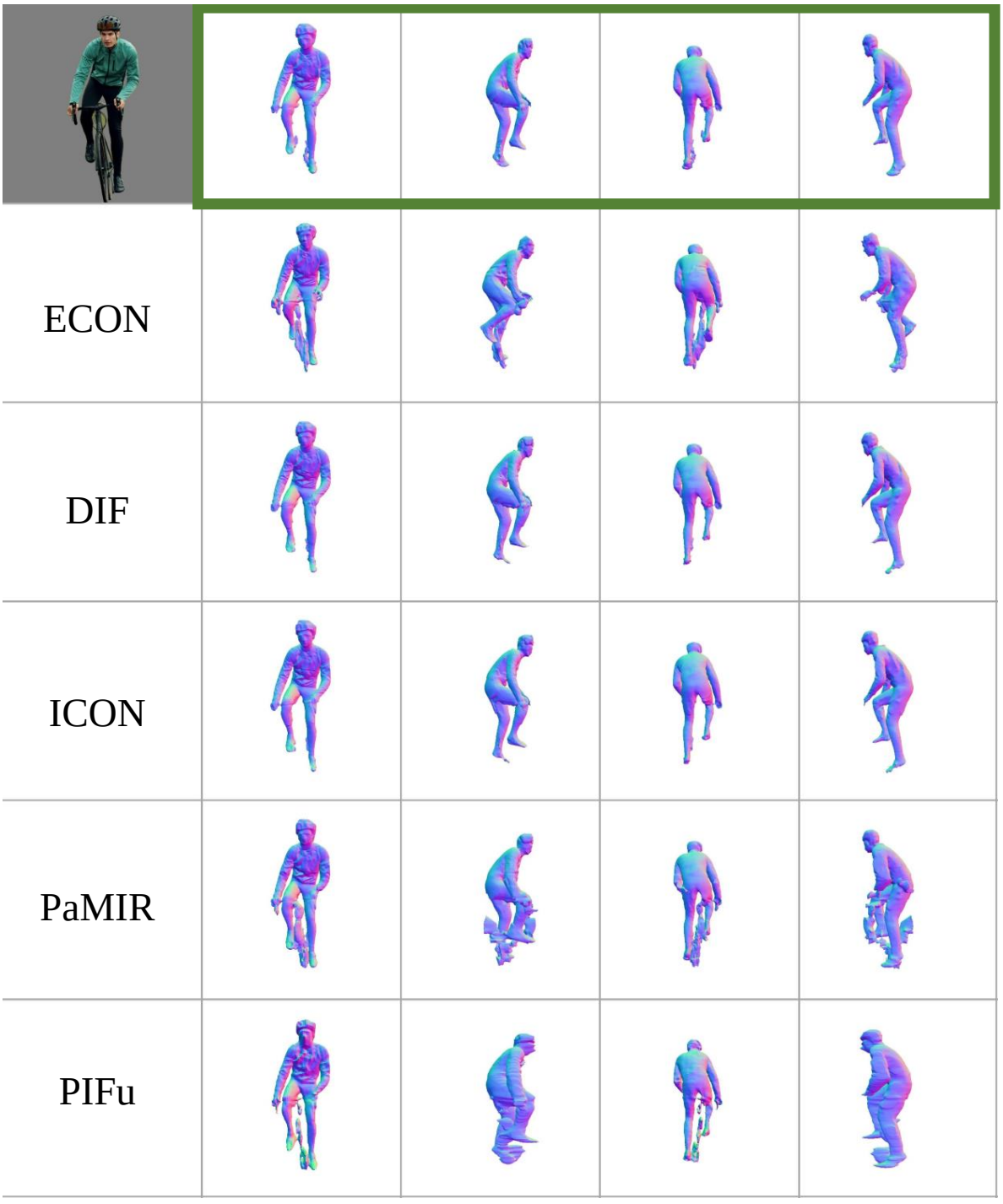

Figure E. Visualization comparisons of reconstruction for our HiLo vs SOTA.

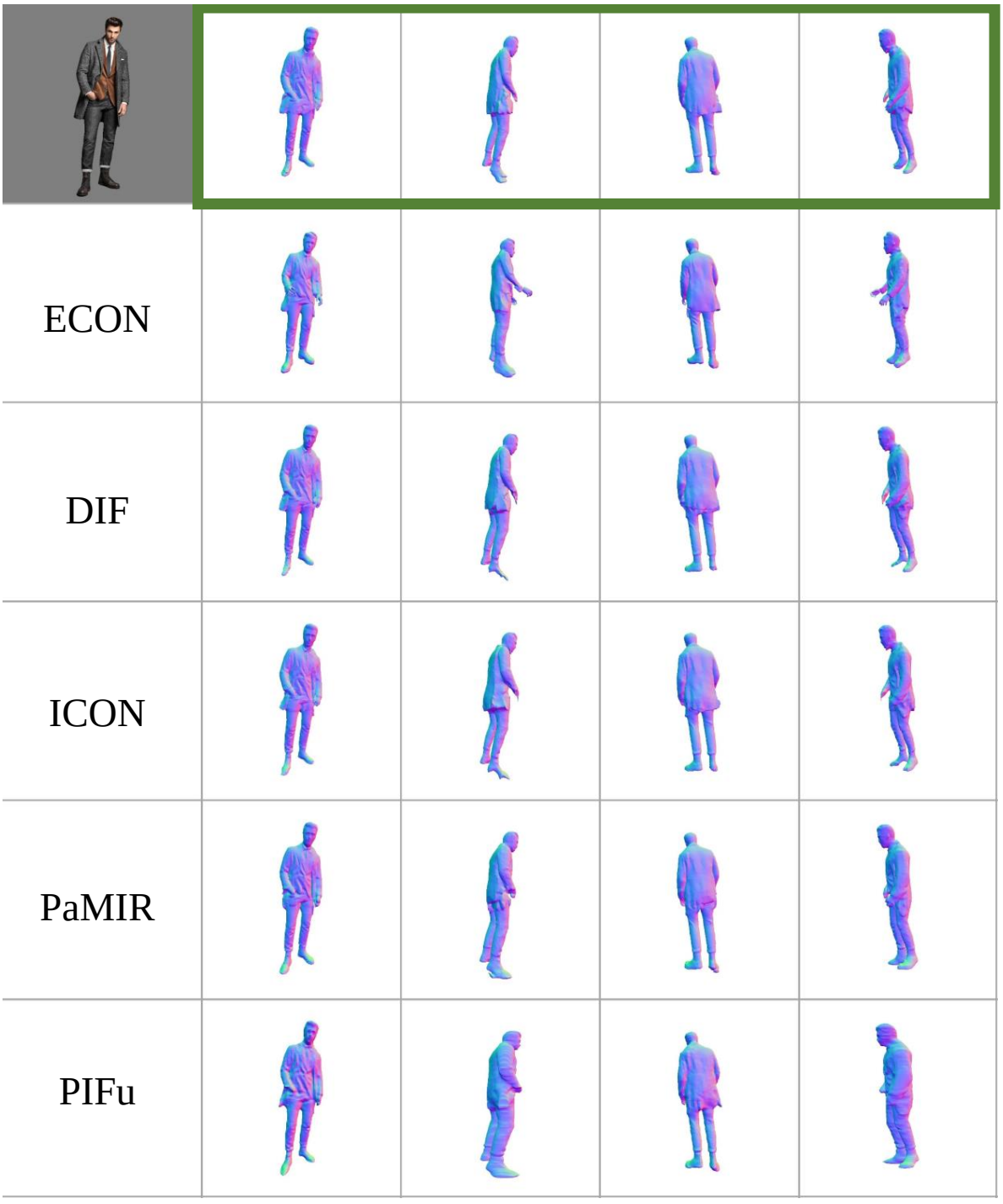

Figure F. Visualization comparisons of reconstruction for our HiLo vs SOTA.

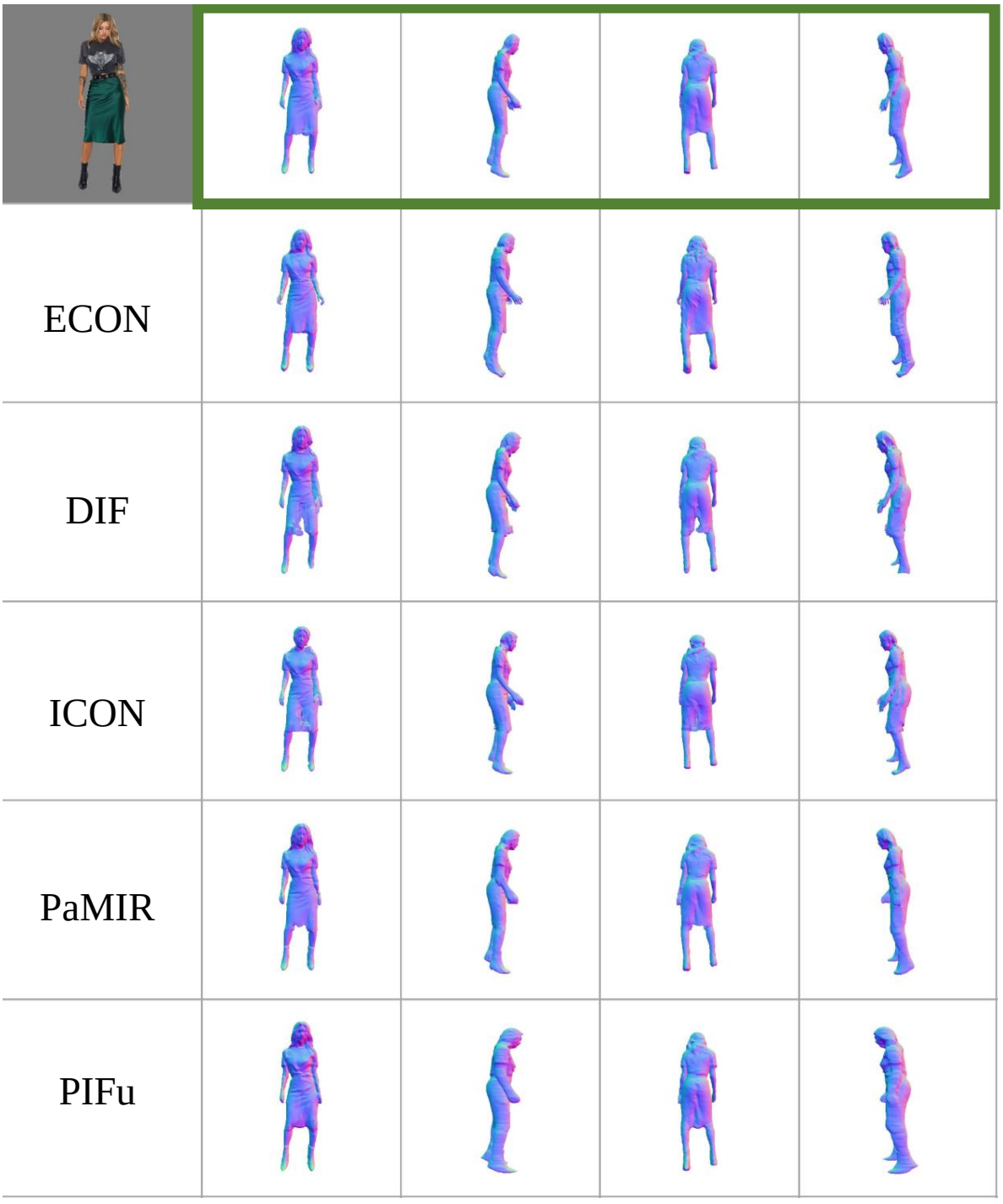

Figure G. Visualization comparisons of reconstruction for our HiLo vs SOTA.

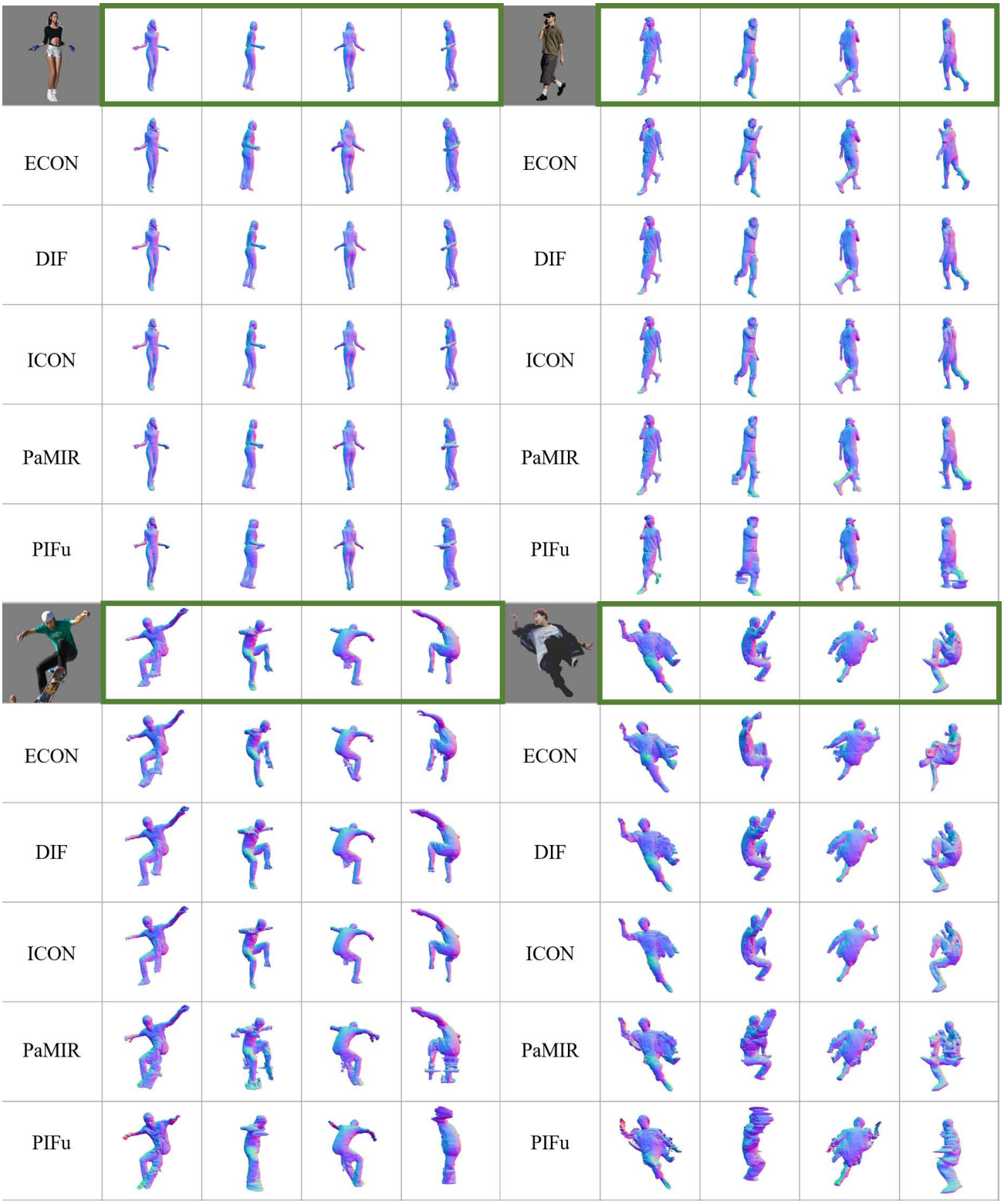

Figure H. Visualization comparisons of reconstruction for our HiLo vs SOTA.

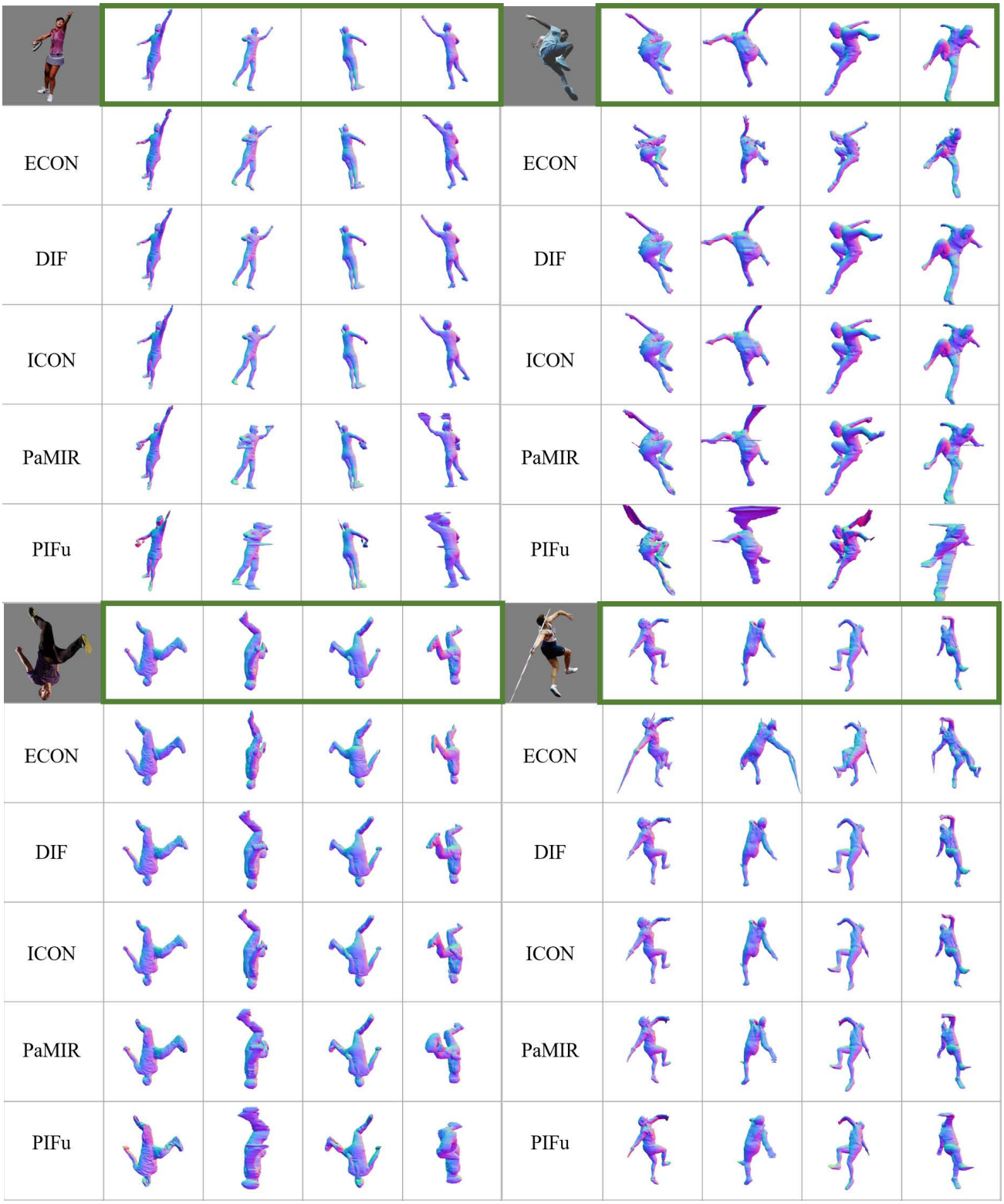

Figure I. Visualization comparisons of reconstruction for our HiLo vs SOTA.

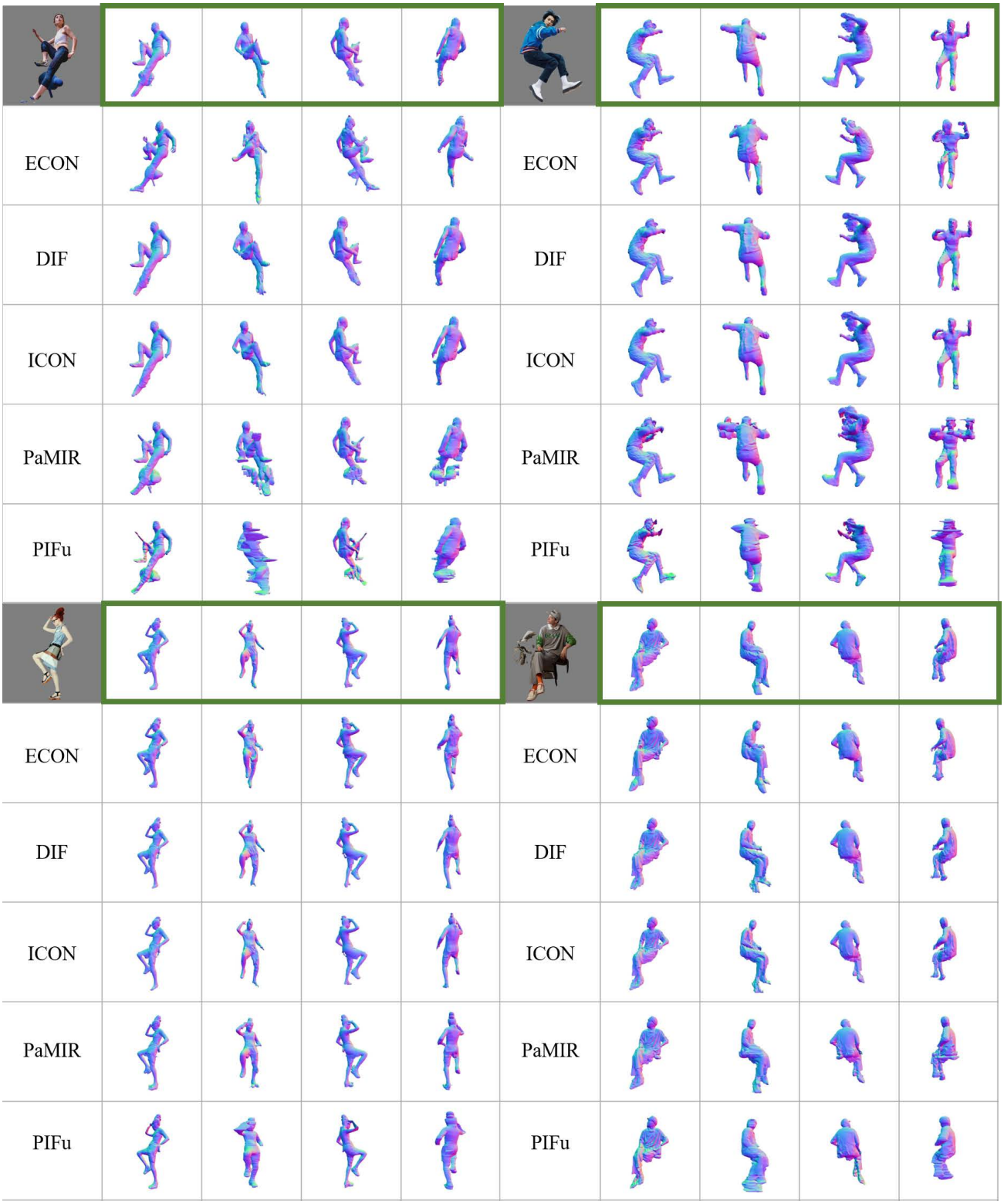

Figure J. Visualization comparisons of reconstruction for our HiLo vs SOTA.

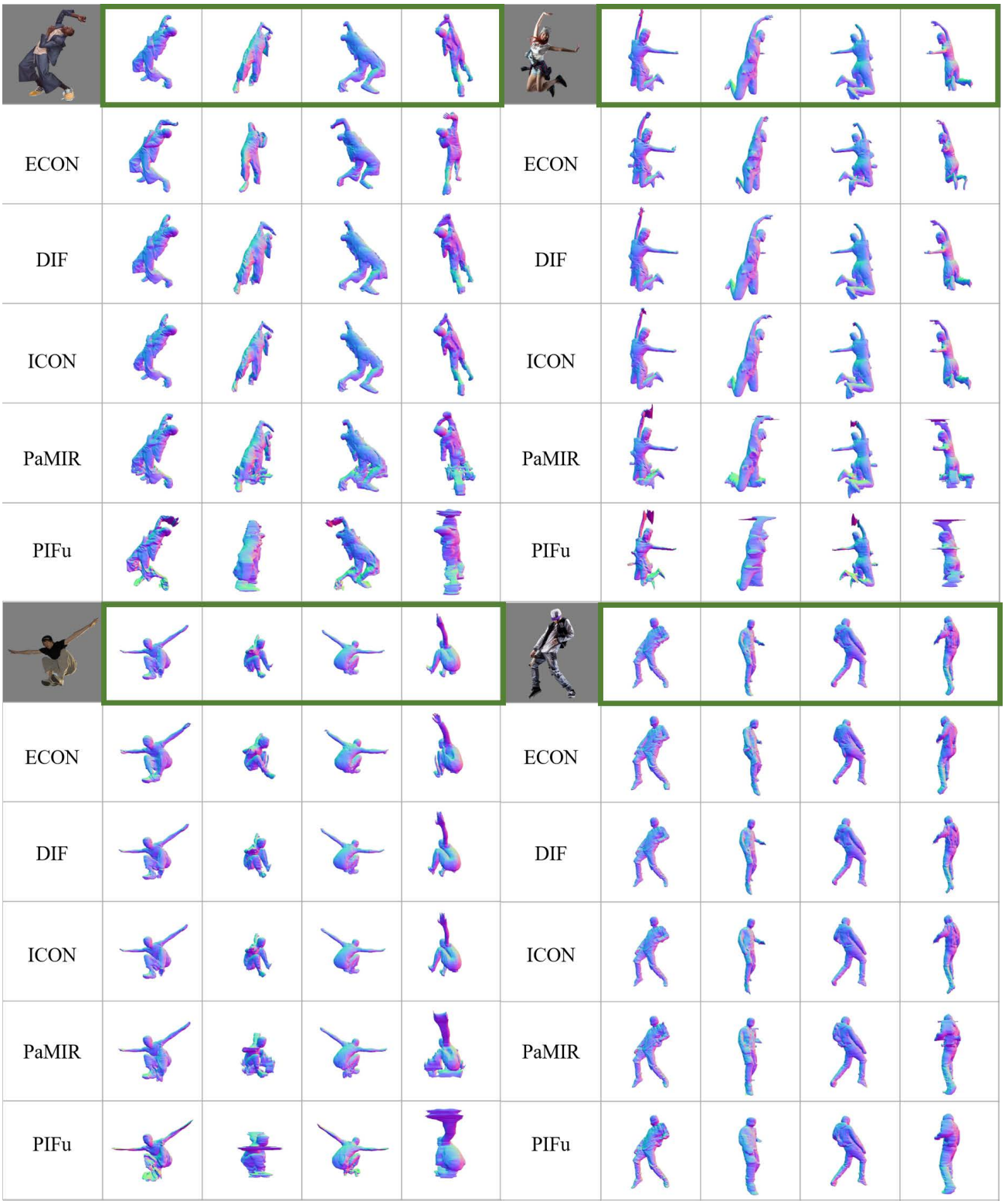

Figure K. Visualization comparisons of reconstruction for our HiLo vs SOTA.