IntegratedPIFu: Integrated Pixel Aligned Implicit Function for Single-view Human Reconstruction

Kennard Yanting Chan[1,2,3], Guosheng Lin[3], Haiyu Zhao[2], and Weisi Lin[1,3]
[1] S-Lab, [2] SenseTime Research, [3] Nanyang Technological University
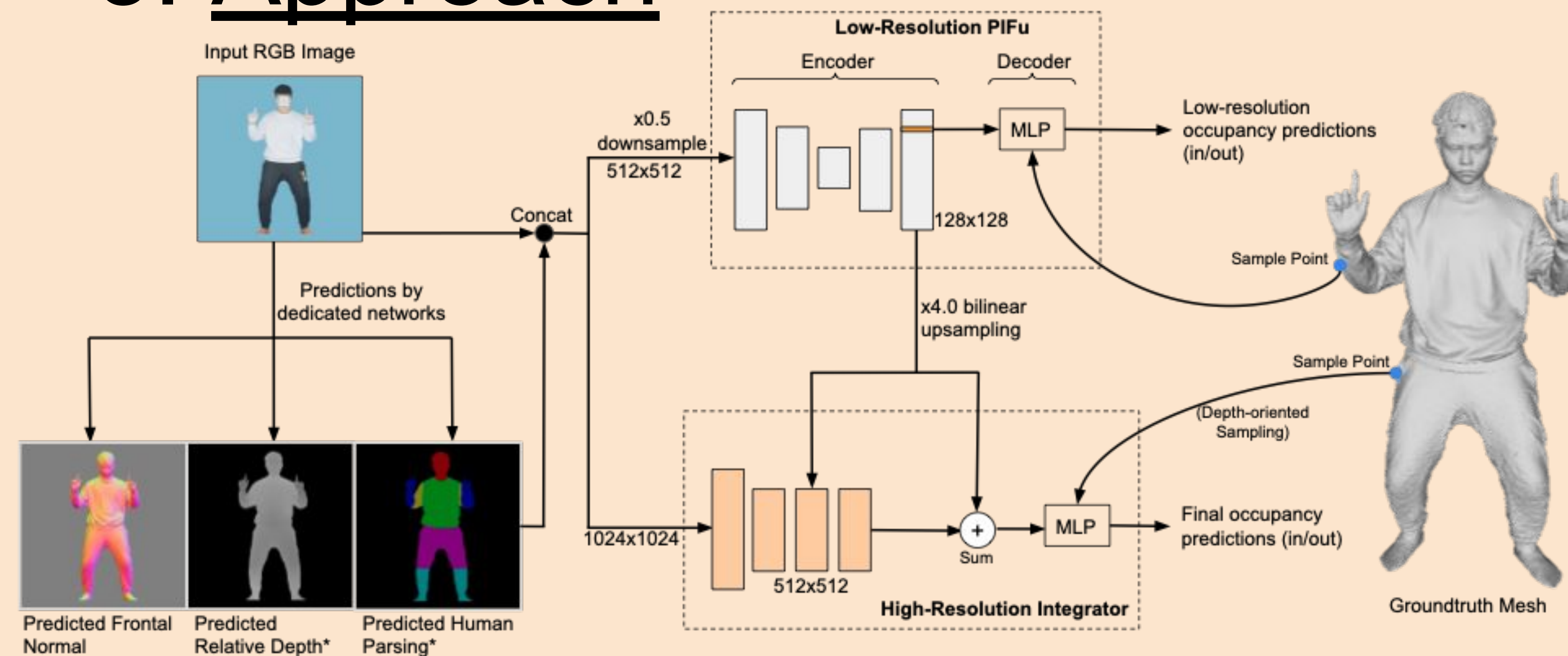
1. Motivation

A well-established class of methods for single-image clothed human reconstruction (SICHR) is pixel-aligned implicit models.

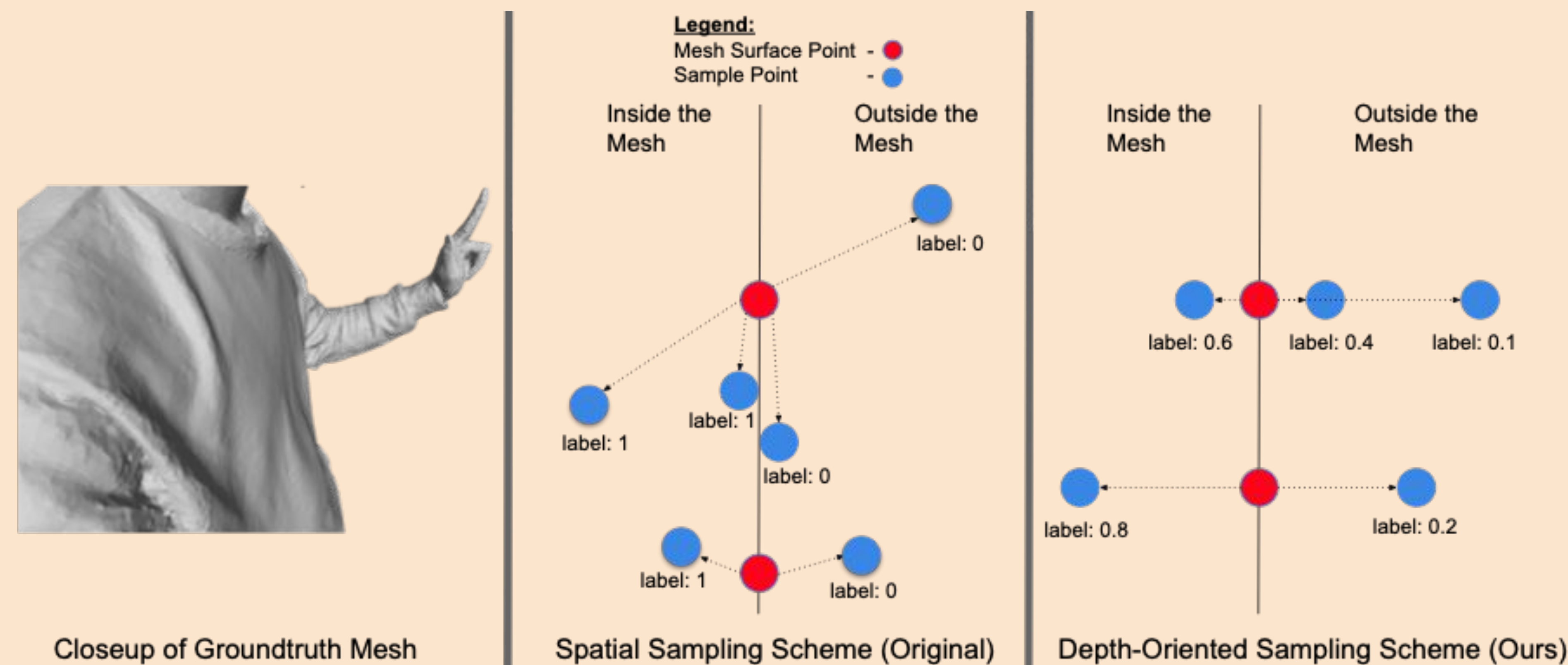But they are prone to problems such as depth ambiguity and generation of meshes with broken limbs.

2. Overview

Thus, we propose IntegratedPIFu, which mitigates the problems by incorporating predicted depth and human parsing maps, depth-oriented sampling (DOS), and High-resolution integrator (HRI).

3. Approach

Input RGB Image

Predictions by dedicated networks

Predicted Frontal Normal

Predicted Relative Depth*

Predicted Human Parsing*

Low-Resolution PIFu

Encoder

Decoder

x0.5 downsample

512x512

128x128

MLP

Low-resolution occupancy predictions (in/out)

Concat

x4.0 bilinear upsampling

Sample Point

1024x1024

512x512

Sum

MLP

(Depth-oriented Sampling)

Sample Point

Final occupancy predictions (in/out)

High-Resolution Integrator

Groundtruth Mesh

An overview of IntegratedPIFu is shown above. Our HRI combines the upsampled feature maps with the 1024x1024 RGB image and predicted frontal normal map. The HRI is then trained using our newly proposed DOS scheme, which is shown below.

Legend:
Mesh Surface Point - ●
Sample Point - ●

Inside the Mesh / Outside the Mesh

label: 0

label: 1

label: 1

label: 1

label: 0

label: 1

label: 0

Inside the Mesh / Outside the Mesh

label: 0.6 label: 0.4 label: 0.1

label: 0.8 label: 0.2

Closeup of Groundtruth Mesh

Spatial Sampling Scheme (Original)

Depth-Oriented Sampling Scheme (Ours)

4. Results and Impacts

(a) Groundtruth   (b) Geo-PIFu   (c) PIFu   (d) PIFuHD   (e) Ours (Depth + Parse)
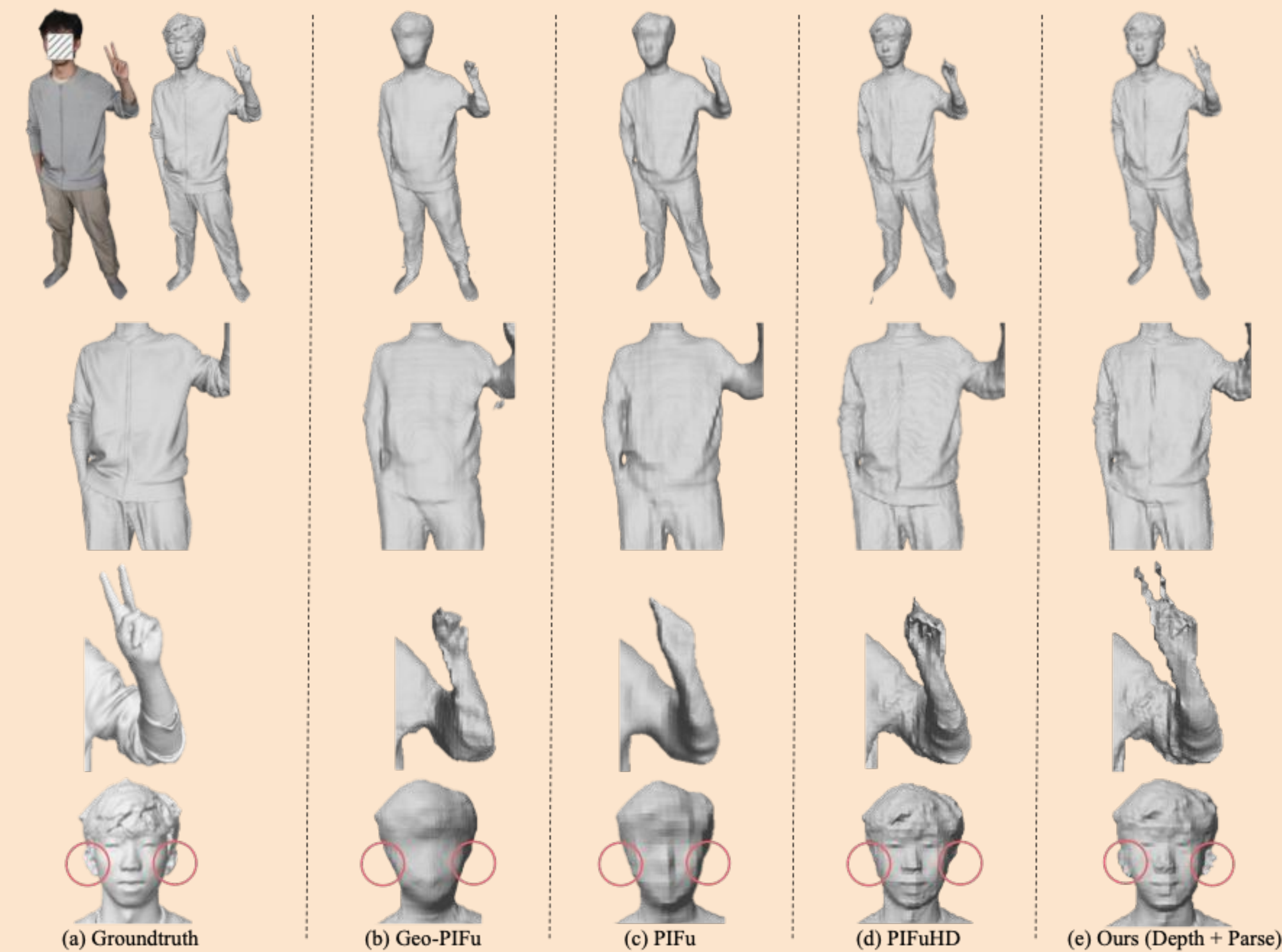
Table 1. Quantitative evaluation of our methods against the state-of-the-arts in the THuman2.0 test set and BUFF dataset. (HRI=High-resolution Integrator, DOS=Depth-Oriented Sampling, D=depth, P=human parsing)

Methods	THuman2.0			BUFF		
	CD (10^{-4})	P2S (10^{-4})	Normal	CD (10^3)	P2S (10^3)	Normal
Geo-PIFu	5.816	9.657	2.452	6.250	10.757	2.912
PIFu	3.135	3.072	1.731	2.639	3.367	1.947
PIFuHD	2.800	2.540	1.698	2.031	2.029	2.010
Ours- HRI	2.841	2.177	**1.622**	2.083	1.931	**1.762**
Ours- HRI+DOS	**2.711**	**2.139**	1.643	**2.029**	**1.925**	1.800
Ours- HRI+DOS (D+P)	3.040	2.259	1.641	2.287	2.224	1.842
Ours- HRI+DOS (D)	3.003	2.387	**1.633**	2.182	2.094	1.858
Ours- HRI+DOS (P)	**2.734**	**2.153**	1.780	**2.030**	**1.897**	**1.789**

1. Our DOS captures delicate human features (e.g. fingers, ears) and fixes th wavy high-frequency artefacts present i PIFuHD (SOTA).
2. Our HRI improves the fidelity of the reconstructed meshes despite using 39% less model parameters than its direct predecessor.