High-fidelity 3D Human Digitization from Single 2K Resolution Images

Sang-Hun Han[1], Min-Gyu Park[2], Ju Hong Yoon[2],
Ju-Mi Kang[2], Young-Jae Park[1] and Hae-Gon Jeon[1]

[1]Gwangju Institute of Science and Technology (GIST), [2]Korea Electronics Technology Institute (KETI)

{sanghunhan, youngjae.park}@gm.gist.ac.kr, haegonj@gist.ac.kr

{mpark, jhyoon, yypeip}@keti.re.kr

Figure 1. Frame-by-frame reconstruction results for a 2K resolution image sequence. The image sequence is captured by a smartphone.

Abstract

*High-quality 3D human body reconstruction requires high-fidelity and large-scale training data and appropriate network design that effectively exploits the high-resolution input images. To tackle these problems, we propose a simple yet effective 3D human digitization method called **2K2K**, which constructs a large-scale 2K human dataset and infers 3D human models from 2K resolution images. The proposed method separately recovers the global shape of a human and its details. The low-resolution depth network predicts the global structure from a low-resolution image, and the part-wise image-to-normal network predicts the details of the 3D human body structure. The high-resolution depth network merges the global 3D shape and the detailed structures to infer the high-resolution front and back side depth maps. Finally, an off-the-shelf mesh generator reconstructs the full 3D human model, which are available at* https://github.com/SangHunHan92/2K2K. *In addition, we also provide 2,050 3D human models, including texture maps, 3D joints, and SMPL parameters for research purposes. In experiments, we demonstrate competitive performance over the recent works on various datasets.*

1. Introduction

Reconstructing photo-realistic 3D human models is one of the actively researched topics in computer vision and graphics. Conventional approaches search for correspondences across multiple views. Therefore, it was necessary to employ multiple camera systems [9, 11] to acquire high-quality human models. However, the bulky and expensive camera systems limit the usage of normal users such as personal content creators and influencers. Recent progress in deep learning has shown the possibility of reconstructing human models from a single image [1, 15, 21, 30, 35, 40, 46, 48, 54, 55]. Nevertheless, there still exists room to improve the quality of 3D human models, especially given an input single image.

Existing approaches fall into two categories; the first is to predict a deep implicit volume [15, 40, 41] and the second is to infer multiple depth maps [13] from an image. In the case of the first approach, the implicit volume can be directly predicted through a deep learning network [46] or the volume can be constructed by predicting each voxel [40, 41] millions of times. Therefore, these approaches are demanding either in memory or in time. The second approach requires to predict at least two depth maps, one for the front and the other for the back, to build a complete 3D human model. Gabeur *et al*. [13] propose an adversarial framework to predict double-sided depth maps; however, it shows poor results owing to inadequate training data generated by using 19 human scan models in addition to the synthetic dataset [47]. Here, researchers have paid attention to recover not only geometric details [41] such as facial regions but also people in various postures by predicting surface normal maps and parametric template models [33].

We claim that the prediction quality of 3D human models is primarily affected by suitably designed deep neu-

ral networks and high-quality training datasets, *i.e.*, high-resolution images and 3D human models. Existing approaches, however, could not handle high-resolution images, *e.g.*, 2048×2048, because it requires massive learnable parameters to train, and there was no such large-scale dataset for 3D human digitization.

There are human datasets opened publicly for research purposes [49–51, 55]. These datasets, however, lack both quality and quantity to train a network generating high-fidelity human models. In this paper, we present a practical approach to reconstructing high-quality human models from high-resolution images and a large-scale human scan dataset consisting of more than 2,000 human models, where existing methods utilize a few hundred human scans to train their networks.

Our framework takes high-resolution images as input up to 2K, 2048×2048, and it is the first to predict high-resolution depth maps for the task of 3D human reconstruction, named **2K2K**. To minimize the number of learnable parameters and memory usage, we split the human body into multiple body parts such as arms, legs, feet, head, and torso, with the aid of a 2D pose human detector [8]. In addition, we align each body part by rotating and scaling to a canonical position which makes the proposed method robust under human pose variation while excluding background regions from the computation. By doing this, the part-wise image-to-normal prediction network can predict accurate surface normals even in the presence of pose variations. Afterward, we merge predicted normal maps into a single normal map and feed it to the normal-to-depth prediction network. Note that it is hard to predict depth maps directly for each body part because of the scale ambiguity; predicting the depth map from a merged normal map can alleviate this problem. We also predict a coarse depth map and feed it to the normal-to-depth prediction network to obtain consistent depth maps over different body parts. Finally, we generate high-fidelity human meshes through Marching cubes [26], whose example is shown in Fig. 1.

To summarize, the contributions of this paper are:

1. **Accuracy.** Our method recovers the details of a human from high-resolution images up to a resolution of 2048×2048.
2. **Efficiency.** The part-wise normal prediction scheme naturally excludes background regions from computation, which results in a reduced memory footprint.
3. **Data.** We release a large-scale 3D human dataset consisting of 2,050 human models with synthesized images.

2. Related Work

We review clothed human 3D reconstruction from single images with respect to datasets, model-free approaches, and parametric model-based approaches. Here, we refer to

Dataset	License	# of subjects	# of Frames	# or vertices	RGB	Mesh	Texture	SMPL(-X)	Keypoint3d
BUFF [51]	Non-commercial	6	>13.6k	150k	✓	✓	✓	✓	✓
CAPE [28]	Non-commercial	15	>140k	6890		✓		✓	✓
DFAUST [5]	Non-commercial	10	>40k	190k		✓		✓	✓
Humbi [50]	Non-commercial	772	26M	50k	✓	✓	✓	✓	✓
THuman [55]	Non-commercial	200	6000	150k	✓	✓	✓	✓	
THuman2.0 [49]	Non-commercial	200	526	300k		✓	✓	✓	
MultiHuman [53]	Non-commercial	150	453	500k	✓	✓		✓	
HuMMan [6]	Non-commercial	1000	60M	300k	✓	✓	✓	✓	✓
RenderPeople [39]	Commercial	-	-	223k	✓	✓	✓		
AXYZ [4]	Commercial	-	-	100k	✓	✓	✓		
Twindom [44]	Commercial	-	-	200k	✓	✓	✓		
2K2K (Ours)	Non-commercial	2050	2050	1M	✓	✓	✓	✓	✓

Table 1. A summary of the clothed human body dataset.

the parametric model as the Skinned Multi-Person Linear (SMPL) model [25, 33] that has been widely used for human pose estimation and reconstruction.

2.1. Datasets for dressed human reconstruction

The clothed human dataset is fundamental for the study of reconstructing the 3D human surface. Early studies use a limited amount of data [5, 51] consisting of few subjects or simple posture. Subsequently, as the parametric model gain popularity in this field, a dataset [28] expressing cloth by modifying SMPL vertices appeared. However, it become necessary to reconstruct human surfaces with more complex postures and difficult clothes. Therefore, a number of studies [49, 53, 55] capture their own datasets and release the data to the public.

On the other hand, commercial human datasets [4, 39, 44] are used for research to ensure professional quality. However, these datasets cost a lot for research. Recently, large-scale human datasets [6, 50] are released but they lack geometry information for detailed face and cloth reconstruction. To overcome this, we release a non-commercial 3D human dataset with the largest number of subjects, *i.e.* humans, and high-resolution geometry features. The characteristics of each dataset are summarized in Table. 1.

2.2. Model-free human reconstruction

BodyNet [46] and DeepHuman [55] exploit human body segmentation or 2D poses and combine them to predict deep implicit volumes. However, the end-to-end training of a volume prediction network requires a large memory footprint and computation. For this reason, BodyNet represents the 3D shape of the human body with a voxel volume whose resolution is $128 \times 128 \times 128$. Caliskan *et al.* [7] use an additional implicit 3D reconstruction module to overcome the low resolution. As one of the model-free approaches, depth prediction [20, 43] is an effective approach to building 3D human models. For example, Mustafa *et al.* [29] show that multiple humans can be reconstructed with the aid of instance segmentation. ARCH [18] and ARCH++ [16] recover the body geometry in a canonical space not only to predict accurate human models but also to animate reconstructed human models. PIFu [40] and its variants [15, 17] train a deep implicit function defined over a pre-defined

(a) Capture studio   (b) 3D model   (c) Rendered image   (d) Texture map   (e) 3D pose   (f) SMPL
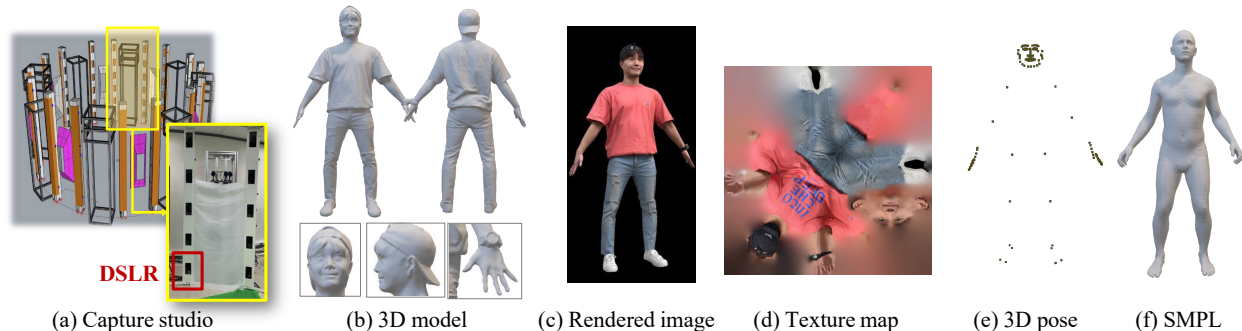
Figure 2. Our datasets provides high-fidelity 3D human models, captured by 80 DSLR cameras, texture maps, 3D poses (openpifpaf-wholebody [24]), and SMPL model parameters [25].
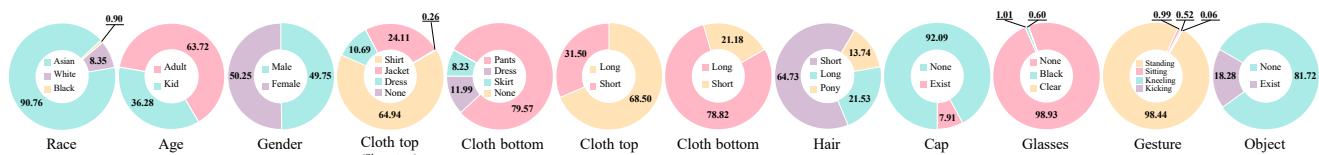

Figure 3. Statistics of the scanned human models in the proposed dataset.

space. Once the implicit function is trained, it is possible to construct an implicit volume by using the implicit function repeatedly, *i.e.*, for each voxel. PHORHUM [3] predicts both geometric and photometric information simultaneously by using global illumination and surface normal prediction. PI-FuHD [41] presents the implicit function for high-resolution images whose input size is up to 1024×1024. They generate high-and low-resolution feature maps and embed them to compute the 3D occupancy field for reconstructing a high-fidelity 3D human model accurately. However, PIFuHD still needs much memory and has slow inference time due to the per-pixel calculation to construct coarse and fine implicit volumes.

Our body part-wise approach allows fast, lightweight, and high-fidelity 3D reconstruction for human models against existing model-free approaches. In addition, we alleviate the potential problems of the model-free approach such as robustness under pose variations.

2.3. Parametric model-based human reconstruction

Pavlakos *et al*. [33] introduce SMPL-X, which proposes latent pose features to exclude invalid human poses efficiently. SMPLpix [36] estimates realistic images with various viewpoints and poses from sparse 3D vertices using RGB-D images obtained from the SMPL model. PaMIR [54] fuses the SMPL model features and image features in the implicit field to reconstruct the detailed 3D human model. Moreover, thanks to the initial geometry cue from the SMPL, it can robustly deal with depth ambiguity. Tan *et al*. [42] propose a supervised approach to predict depth maps from a video by imposing the consistency between different frames through the initial SMPL model. SMPLicit [10] predicts the unsigned distance to the object surface using the SMPL oc-

clusion map in the latent space. Zhu *et al*. [56,57] propose to deform initial SMPL vertices using image-aligned features. ICON [48] reconstructs clothed human surfaces in various postures by feeding initial normal maps from the predicted SMPL model. DensePose [14] introduces a simplified human mesh model that is represented by the predefined UV map. Alldieck *et al*. [2] utilize DensePose to extract and project human image textures in the UV map. Their method first infers surface normal maps from the UV texture map and afterward reconstructs a 3D human model in a single predefined pose, *e.g*., T or A pose. Jafarian and Park [19] adopt DensePose to infer depth information from the video. Having the depth information of a previous frame in one hand, they sequentially improve the details of 3D human body structures. Zhao *et al*. [52] modify the SMPL templates for A-posed humans and compute the dynamic offset through the UV map to reconstruct human models.

Unlike the above works, our 2K2K only needs single images to capture fine 3D details. In addition, the proposed method can deal with various human poses, clothes, and accessories because it does not count on the parametric human model, which has difficulty capturing geometric details.

3. Human Scan Dataset

Since the performance of human reconstruction primarily depends on the quality and quantity of training data, we release our high-quality human scan models in public, consisting of 2,050 3D human models scanned by our scan booth. The booth contains 80 DSLR cameras, *i.e*., Nikon D50 model, where the cameras are aligned vertically and horizontally as illustrated in Fig. 2 (a). The pitch angle of a camera is adjusted manually, assuming an average human height is standing in the middle, *i.e*., 5.9 feet tall. The highest
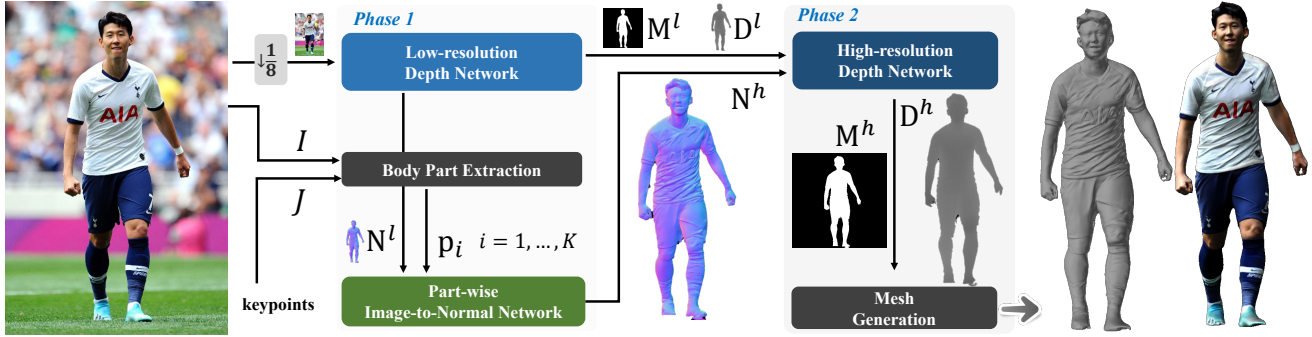
Figure 4. An overall framework of the proposed method. The first phase predicts the low-resolution front/back view depth maps, and the high-resolution front/back view normal maps. The high-resolution depth network upsamples the low-resolution depth maps with the guidance of the high-resolution normal maps. Finally, the mesh generation reconstructs the full 3D model.

cameras look at the head region; the second-highest cameras look at the upper body; the third-highest cameras capture the mid-body region; the fourth-highest cameras look at the lower body; the lowest cameras are pointing at the knees. We generate initial human models by using commercial software, i.e., RealityCapture [38], and manually post-process the models by filling holes and enhancing geometric details in hair regions by experts. Finally, we decimate meshes while keeping important geometric structures. The number of vertices for released models is about 1M, whose example is shown in Fig. 2 (b). The scanned model preserves the geometric details such as fingers and wrinkles properly mainly because of high-quality images captured in a controlled environment.

3.1. Synthesizing images for training

Instead of directly using images taken in our scan booth, we augment our data to mimic realistic environments and diversify the training data. There are many ways to render such images [47]. In terms of synthesizing approaches, it is possible to generate photo-realistic images with environmental assets and rendering software such as Unreal [12] and Unity [45]. An example result is shown in Fig. 2(c), where the reflecting materials and shadings are rendered realistically. However, this ray-tracing approach is time-consuming and requires manual settings by experts. Therefore, we use a more straightforward approach i.e., the Phong lighting model while superimposing the background using randomly selected images. Therefore, one can also recover shadings and lighting directions from the shaded input. In addition, we also note that multiple humans can be placed at random positions, or a single person can be placed at the center of the image. The latter assumes that humans are already detected and appropriately cropped, thanks to the recent advances in object detection technology. We additionally provide texture map, 3D keypoint, and SMPL Model as shown in Fig. 2(d),(e),(f) for the effectiveness of the dataset. In this paper, we generate human-centered training images, whose details will be described in Sec.5.

3.2. Statistics and bias

There are many kinds of human appearances in the real world, determined by various factors such as gender, country, and culture. We believe, therefore, it is crucial to share statistics of scanned humans to make our dataset applicable to diverse research topics. In order to compile distribution statistics for our dataset, we limit those to 12 categories as shown in Fig. 3. Due to the geographic circumstances, more than 90% of models are from Asian countries, which is a common issue on the person-related dataset. By considering various combinations of the treat, our dataset enables deep learning networks to form diverse feature representations of human bodies.

4. Proposed Method

We aim to produce a high-quality 3D human model $\mathbf{V}$ from a single high-resolution image I with a resolution of up to 2048×2048. As illustrated in Fig. 4, the proposed method $\Phi(\cdot)$ conducts reconstruction formulated as below:

$$\mathbf{V} = \Phi(I, \mathcal{J}, \mathcal{S}), \tag{1}$$

where $j_i \in \mathcal{J}$ refers an ith joint and $s_i \in \mathcal{S}$ denotes predefined patch size cropped centered at j_i. Using these inputs, the first phase predicts low-resolution depth maps $\mathbf{D}^l$ and high-resolution normal maps $\mathbf{N}^h$ from low-resolution normal maps $\mathbf{N}^l$. The second phase produces the high-resolution front and back-sided depth maps $\mathbf{D}^h$ from $\mathbf{D}^l$ and $\mathbf{N}^h$. Finally, the high-fidelity 3D model $\mathbf{V}$ is reconstructed from $\mathbf{D}^h$.

4.1. Part-wise image-to-normal network

Body part extraction and alignment. Since the image-to-normal network takes aligned part images as input, we first crop image patches using 2D joints for the 12 body parts i.e., head, torso, upper arms, lower arms, upper legs, lower legs, and feet. Let $\mathcal{P} = \{\mathbf{p}_i | 0 \leq i < 12\}$ be the twelve patches cropped from image I by using 2D joints $\mathcal{J}$, centered at 2

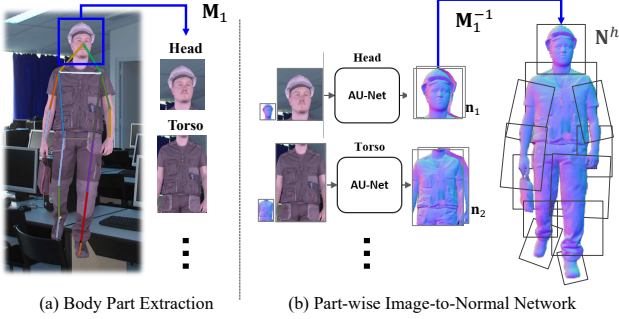(a) Body Part Extraction   (b) Part-wise Image-to-Normal Network

Figure 5. (a) Based on human joints, the body parts are categorized into five parts, *i.e.* a head, a torso, arms, legs, and feet. Afterward, we transform them in the predefined sizes and positions and align the same body parts in the same direction to have a roughly similar structure. (b) We adopt the AU-Net [31] to generate front and back view part-wise normal maps $\mathbf{n}_i$. Then, we warp each normal map to the original position and scale with the transformation matrix $\mathbf{M}_i$, and then finally obtain the high-resolution normal maps $\mathbf{N}^h$.

to 4 joints. For example, the upper arm region is extracted by using two joints, the shoulder and elbow, and the torso region is extracted by using two shoulder joints and two pelvis joints. Then, we warp image patches,

$$\bar{\mathbf{p}}_i = \mathbf{M}_i \mathbf{p}_i, \qquad (2)$$

where $\bar{\mathbf{p}}_i \in \bar{\mathcal{P}}$ is the body part transformed by a similarity transformation $\mathbf{M}_i$. An example of cropped images is illustrated in Fig. 5 (a), where we feed $\bar{\mathcal{P}}$ to the normal-to-image network. An inverse similarity transformation $\mathbf{M}_i^{-1}$ is then used for merging the part-wise normal maps by inversely warping each patch. This pipeline is simple yet effective to train high-resolution images not only because it reduces the computational complexity but also because a less complicated network can be used for aligned part-wise depth prediction.

Part-wise normal prediction. We generate the part-wise surface normals by using an image-to-normal network $G_{\mathbf{N},i}(\cdot)$ as below:

$$\bar{\mathbf{n}}_i = G_{\mathbf{N},i}(\bar{\mathbf{p}}_i, \mathbf{M}_i^{-1}\mathbf{N}^l), \qquad (3)$$

where $\bar{\mathbf{n}}_i$ denotes a double-sided normal map for i^{th} aligned body part. The double-sided normal map indicates the stacked normal maps for front and rear views. The $\mathbf{N}^l$ is used to separate the background and provide an initial geometry, which will be explained in Sec. 4.2. Note that instead of training twelve independent networks, we train the five image-to-normal prediction networks for head, torso, arms, legs, and foot regions, respectively, as shown in Fig. 5 (b). The separation scheme enforces each network to focus on its unique structure and patterns of the body parts as well as to reduce the number of learnable parameters. In addition, we use the AU-Net [31] as our backbone network to predict
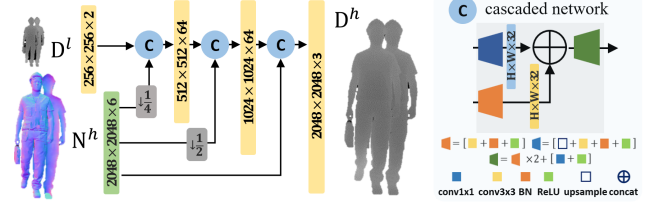

Figure 6. The high-resolution depth network takes the low-resolution depth and the high-resolution normal maps in Eq. (4) to render the high-resolution depth maps $\mathbf{D}^h$. All of normal maps and depth maps consist their front and back views.

normal maps. Once normal maps are predicted for all body parts, we assemble them into the high-resolution normal maps $\mathbf{N}^h$ as follows:

$$\mathbf{N}^h = \sum_{i=1}^{K} \left(\mathbf{W}_i \odot \mathbf{n}_i \right), \qquad (4)$$

where $\mathbf{n}_i = \mathbf{M}_i^{-1}\bar{\mathbf{n}}_i$ is an inverse warped normal map of $\bar{\mathbf{n}}_i$ by applying $\mathbf{M}_i^{-1}$ and $\odot$ stands for an element-wise product. Since the multiple normal vectors can be projected onto the same pixel in the original image, we blend normal vectors when they are near the boundary region and projected multiple times. The blending weight $\mathbf{W}_i(x, y)$ is computed as,

$$
\mathbf{W}_i(x, y) = \frac{G(x, y) * \phi_i(x, y)}{\sum_i G(x, y) * \phi_i(x, y)},
$$
$$
\phi_i(x, y) = \begin{cases} 1 & \text{if } \sum \mathbf{n}_i(x, y) \neq \mathbf{0}^\top \\ 0 & \text{otherwise} \end{cases}, \qquad (5)
$$

where (x, y) denotes a pixel location, $*$ stands for a convolution operator, and $\mathbf{0}^\top$ is the six dimensional zero vector. We compute a pixel weight $\mathbf{W}_i(x, y)$ w.r.t. the pixels having projected normal vectors which results in decreasing weights near the boundary regions conditioned by the size of a Gaussian kernel $G(\cdot, \cdot)$ with the boundary region $\sigma = 7px$.

4.2. Low-resolution depth prediction network

Rather than directly predicting the depth maps from the merged normal maps $\mathbf{N}^h$, we predict the coarse depth maps to guide predicting high-resolution depth maps. The low-resolution depth network sequentially infers coarse normal maps $\mathbf{N}^l$ and depth maps $\mathbf{D}^l$. Although it lacks geometric details, the smaller image size better captures the global geometric information. In this work, we adopt the depth prediction network used in [32] because the dual-encoder AU-Net (D-AU-Net) selectively captures meaningful features from photometric and geometric information, therefore, different information can be complementary each other. On the other hand, the attention module makes the network converges quickly compared to the conventional U-Net structure. We first down-sample the input image I to a low-resolution

image I^l. The image-to-normal network in the D-AU-Net predicts double-sided coarse normal maps $\mathbf{N}^l$ from $\mathbf{I}^l$. Different from [32] which generates double-sided depth and normal maps, our low-resolution depth network $G_{\mathbf{D}^l(\cdot)}$ additionally yields a mask map $\hat{\mathbf{M}}^l$ to deal with the background. Finally, the low-resolution depth map $\mathbf{D}^l$ is computed by the element-wise product $\odot$ of $\hat{\mathbf{D}}^l$ and $\hat{\mathbf{M}}^l$.

$$\begin{aligned} \mathbf{D}^l &= \hat{\mathbf{D}}^l \odot \hat{\mathbf{M}}^l, \\ \hat{\mathbf{D}}^l, \hat{\mathbf{M}}^l, \mathbf{N}^l &= G_{\mathbf{D}}^l(I^l). \end{aligned} \qquad (6)$$

Here, we set the resolution of an input image to 256×256 which is smaller than the input by a factor of 8.

4.3. High-resolution depth prediction network

We design the high-resolution depth network based on a shallow convolution network architecture that consists of three cascaded blocks as shown in Fig. 6. The high-resolution depth network $G_{\mathbf{D}}^h(\cdot)$ is defined as below:

$$\begin{aligned} \mathbf{D}^h &= \hat{\mathbf{D}}^h \odot \hat{\mathbf{M}}^h, \\ \hat{\mathbf{D}}^h, \hat{\mathbf{M}}^h &= G_{\mathbf{D}}^h(\mathbf{N}^h, \mathbf{D}^l), \end{aligned} \qquad (7)$$

where it takes the low-resolution depth maps $\mathbf{D}^l$ and the high-resolution surface normal maps $\mathbf{N}^h$ as input to predict high-fidelity depth maps $\mathbf{D}^h$ via the element-wise product $\odot$ of $\hat{\mathbf{D}}^h$ and $\hat{\mathbf{M}}^h$. The cascade block basically increase the resolution of input by the factor of 2, as illustrated in Fig. 6. Here, the first two blocks fuse depth and normal information to generate feature maps, and the last block generates double-sided depth map $\mathbf{D}^h$ from $\mathbf{N}^h$ and fused features. Unlike the existing works that calculate the signed distance function through the implicit function of the MLP structure, we choose the CNN structure to encode the normal locality information. This contributes to forming a detailed geometry by exchanging information with adjacent pixels in the network structure.

Mesh generation. There are multiple ways to generate a 3D model from depth maps. In this work. we take a similar approach to [13]. We convert the depth maps into 3D point clouds, and then compute a surface normal for each point from its neighboring points. Afterward, we run a screened Poisson surface construction [22] to obtain the smooth human mesh $\mathbf{V}$.

4.4. Training scheme and loss functions

To train our network, we divide the training procedure into two phases while employing different loss functions. In the *phase1* of Fig. 4, we use both a L_1 loss and L_{SSIM} loss to optimize the image-to-normal networks, $G_{\mathbf{N},i}(\cdot)$ and $G_{\mathbf{N}}^l(\cdot)$ as follows:

$$\begin{aligned} L_{phase1, N} = \alpha L_1(\mathbf{N}_{GT}^l, \mathbf{N}^l) + \beta L_{\text{SSIM}}(\mathbf{N}_{GT}^l, \mathbf{N}^l) + \\ \sum_i(\alpha L_1(\bar{\mathbf{n}}_{GT,i},\ \bar{\mathbf{n}}_i) + \beta L_{\text{SSIM}}(\bar{\mathbf{n}}_{GT,i},\ \bar{\mathbf{n}}_i)), \end{aligned} \quad (8)$$

where L_{SSIM} denotes the SSIM loss to preserve high-frequency information in shape boundaries. The depth network $G_{\mathbf{D}}^l(\cdot)$ is then trained by minimizing a linear combination of a smooth-L_1 loss L_{s1} and binary cross-entropy loss L_{BCE} as below:

$$\begin{aligned} L_{phase1, D} = \alpha L_{s1}(\mathbf{D}_{GT}^l, \mathbf{D}^l) + \\ \beta L_{s1}(\mathbf{N}_{GT}^l, \mathbf{N}_D^l) + \gamma L_{BCE}(\mathbf{M}_{GT}^l, \mathbf{M}_D^l). \end{aligned} \quad (9)$$

We also utilize a regularization term to maintain a consistency between the ground-truth normal maps and the normal maps $\mathbf{N}_D^l$ converted from the predicted depth maps $\mathbf{D}^l$, similar to [19]. L_{BCE} is used to learn the foreground mask with the binary cross-entropy loss. Three hyper-parameters are empirically set to α=0.85, β=0.15, and γ=0.15.

For training the *phase 2* in Fig. 4, we freeze the networks trained in the *phase 1* and train for the high-resolution depth generator $G_{\mathbf{D}}^h(\cdot)$. For this, we use the same losses used in Eq. (9):

$$\begin{aligned} L_{phase2} = \alpha L_{sl1}(\mathbf{D}_{GT}^h, \mathbf{D}^h) + \\ \beta L_{sl1}(\mathbf{N}_{GT}^h, \mathbf{N}_D^h) + \gamma L_{BCE}(\mathbf{M}_{GT}^h, \mathbf{M}_D^h), \end{aligned} \quad (10)$$

We also use the same hyper-parameters as in Eq. (9). Intuitively, fine-tuning the entire models can be a viable option.

5. Experimental Results

For training, we use Adam optimizer [23] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and stochastic gradient descent with the warm restarts scheduler with $T_0 = 5$ [27]. The learning rate and batch size are set to 0.0001 and 2, respectively. We set the kernel size to 15×15 and σ to 5 for the Gaussian smoothing kernel $G(x, y)$. We train our model for 30 epochs using the publicly available PyTorch framework, which takes about 3 days on a machine with four NVIDIA RTX A6000 GPUs.

5.1. Training data generation

Overall, we scanned 2,050 human models. We will release our human scan models to the public upon the acceptance of this paper. We also trained our network by using all the human models to see how far the proposed framework can improve the performance of human model generation. In addition, 368 high-quality human models [39] are employed, denoted as the RenderPeople dataset, and the THuman2.0 [49] model was also used for this experiment. We synthesized a background image from [37] to the input image to increase the reconstruction efficiency for the real environment.

To train our network, images and depth maps are generated for front and back views through perspective projection, and the normal maps are then obtained from the generated depth maps. First, we shift each human model to be located at the center of the rendered image using the centroid value. Then, we rotate the shifted model horizontally from -30 to

Method	RenderPeople				THuman2.0				Inference Time(s)↓
	P2S↓	Chamfer↓	Normal↓	IoU↑	P2S↓	Chamfer↓	Normal↓	IoU↑	
PIFu	1.96	2.10	8.11	64.1	3.10	3.02	8.52	59.5	4.16
PIFuHD	1.45	1.66	6.56	70.1	2.42	2.81	8.89	59.7	13.67
PaMIR	1.51	1.68	6.50	67.8	2.04	2.43	6.19	71.6	2.70
ICON	**0.55**	**0.57**	3.78	**87.5**	**1.21**	**1.19**	6.65	**74.6**	9.61
2K2K	1.12	0.92	**3.64**	84.4	1.53	1.65	**5.71**	69.0	**0.31**

Table 2. Single-view human reconstruction evaluation results on the RenderPeople and THuman2.0 datasets.

30 with an interval of 10 degrees. Here, the virtual camera position faces the subjects from the [0, 0, -1] position. The field of view is 50 degrees, and the image resolution is 2048×2048. In the rendering process, ambient and diffuse lighting is placed around the model to create a realistic training image. The ambient lighting applied to the object's center is fixed, the intensity and color of diffuse lighting are randomly changed, and 90 diffuse lights are placed around the object. As a result, shaded front images, back images, depth maps, and normal maps are generated for training.

In addition, we demonstrate the robustness of our network. To do this, we use two public datasets: RenderPeople [39] and THuman2.0 dataset [49]. We purchase 368 commercial subjects for RenderPeople dataset and split them whose train/test set ratio is about 9:1 (331/37). For THuman2.0, the number of the train and test sets is 500 and 26, respectively, following the evaluation protocol in ICON [48] for fair comparison. As an ablation study, we demonstrate both the usefulness of our dataset and its synergy with our network by checking the performance improvement of our network with respect to the number of training images and its resolution. For the ablation study, we use up to 2,000 training subjects and 50 test subjects.

5.2. Quantitative and qualitative evaluation

We compare 2K2K with the state-of-the-art (SOTA) methods, including PIFu [40], PIFuHD [41], PaMIR [54] and ICON [48] with human scanned data from RenderPeople [39], THuman2.0 [49]. For evaluation, we use popularly used metrics in the literature: average point-to-surface(P2S) distance, Chamfer distance, average surface normal error, and Intersection over Union(IoU). Metric units are centimeters for P2S and Chamfer, and radian for surface normal. For the P2S distance and Chamfer distance metrics, RenderPeople used the original scale, and Thuman2.0 was evaluated by normalizing the mesh scale for fair evaluation because the scale is different for each subject.

Table. 2 shows the quantitative comparison results. Ours achieves the best results in almost measures. Although ICON shows promising result in terms of the P2S distance and Chamfer distance, we outperform the SOTA models in the surface normal metric, thanks to the strength of CNN architecture taking locality information to the end of the network. We add its detailed explanation in supplementary. For inference time, other implicit function methods require a huge

Resolution	P2S↓	Chamfer↓	Normal↓	Inference Time(s)↓	Memory(MB)	
					Whole Img.	Part-wise
512×512	1.17	1.20	4.72	**0.043**	6,565	5,735(87%)
1024×1024	1.15	1.10	4.36	0.091	15,065	11,011(73%)
2048×2048	**0.96**	**1.01**	**3.69**	0.308	47,195	30,413(**64%**)

Table 3. Performance evaluation w.r.t. the different resolution inputs. As the resolution of input becomes larger, our approach more effectively predicts the depth maps. Our method predicts high-quality depth maps at 20/10/3 fps for different resolutions of inputs.

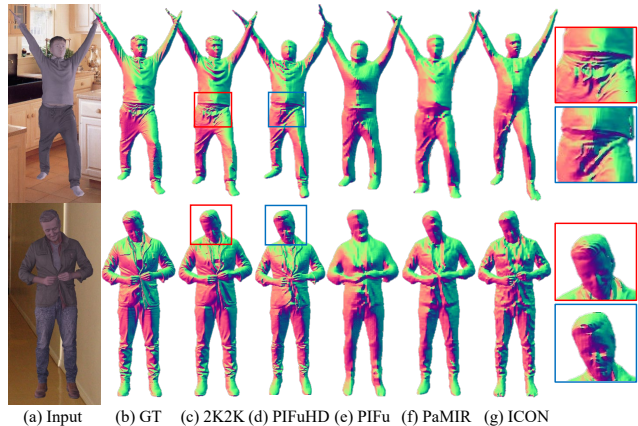(a) Input (b) GT (c) 2K2K (d) PIFuHD (e) PIFu (f) PaMIR (g) ICON

Figure 7. Qualitative results. 2K2K shows accurate results over (c)-(g). Particularly, our results show that the detailed shape of the human body are better expressed compared to other methods.

time to compute the signed distance function, while since our method predicts the depth map in the end-to-end manner, the inference is done within 0.3 seconds even with 2K resolution images. Fig. 7 compares the normal maps by using predicted human models from RenderPeople and THuman2.0 images. As shown, our method can recover details of facial regions compared to the other methods.

Effectiveness. We evaluated the memory footprint, *e.g.*, required memory for a single batch training, at different resolutions. Table. 3 shows the actual memory usage during training and the perfomance changes. When we use 2K resolution images, we can obtain the impressive results as expected. In addition, our part-based approach enable our network to use the high-resolution images as input. That means that the proposed method is more efficient and powerful than those of the direct usage of whole images as input.

Geometric detail. Table. 3 shows the performance according to the image resolutions in our dataset. The P2S and Chamfer metrics imply that there is a little performance gain when using higher-resolution images. However, we observe that the significant improvements exist in terms of the normal metric, which means high-frequency details are well preserved. In addition, Fig. 9 compares the different quality of outputs w.r.t the resolution changes. As expected, 1024×1024 and 512×512 results show blurry results because downsampling the image erases meaningful information. The most important comparison is in Fig. 9 (a) and (b) where the input resolution is same but the whole-body
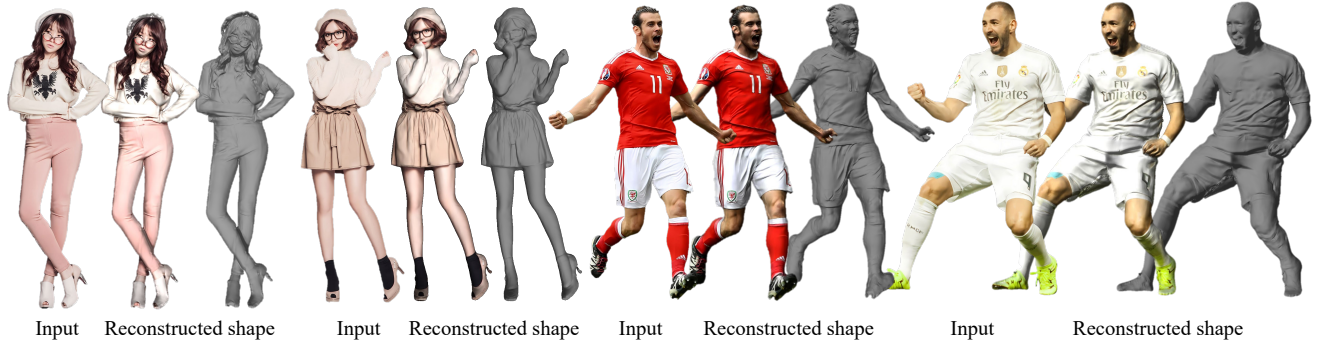
Input   Reconstructed shape   Input   Reconstructed shape   Input   Reconstructed shape   Input   Reconstructed shape

Figure 8. Single image 3D human reconstruction results in the wild. The images are downloaded in internet.

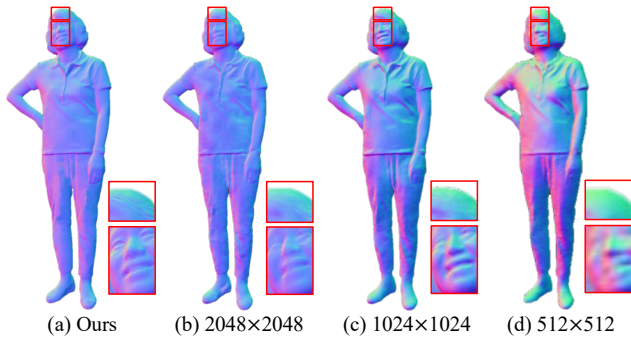(a) Ours   (b) 2048×2048   (c) 1024×1024   (d) 512×512

Figure 9. An example of our result. (b) is the part-ware result of the 2048 resolution image, and (c) (e) are the results of direct prediction of normals in the 2048, 1024, and 512 resolution images.

Figure 10. Smooth side-view quality of the reconstructed model.

approach loses details compared to the part-wise approach. This indicates that the part-wise approach is not only efficient but also accurate, with the aid of the part-wise normal prediction scheme.

Fig. 10 shows the smooth surface of the reconstructed human model even from the side-view. We note that our method works well for non-front-facing humans because we rotate human models in the augmentation step. In addition, Table. 4 shows the performance change according to the number of training datasets. The metric indicates that the overall performance is saturated when the number of the subjects is 2,000.

In-the-wild performance. Fig. 8 shows the qualitative result made from internet photos. Although image and lighting information not included in the training set, our method shows high-fidelity reconstructions from in-the-wild images.

Limitations. Since we explicitly predict normal maps for each body part, our method do not take severe self-occlusion into account, *e.g.*, when a lower arm is behind the back. We claim that this phenomenon is inherently ambiguous, possi-

# of scan models	P2S↓	Chamfer↓	Normal↓
250	1.44	1.35	7.00
500	0.99	1.05	4.36
1,000	0.98	1.04	4.06
2,000	**0.965**	**1.010**	**3.65**

Table 4. A comparison of performance changes according to the number of training datasets. This table verifies that the number of scan models is an important factor.

ble remedies are either predicting semantics for occluding pixels or to employ human body prior [34] to guide the depth prediction. We show several failure cases in the supplementary material due to the space limit.

6. Conclusion

We have proposed 2K2K, an effective framework for digitizing humans from high-resolution single images. To achieve this, we first built a large-scale human model dataset by scanning 2,050 human models, and used them to train our networks, consisting of part-wise normal prediction, low-resolution, and high-resolution depth prediction networks. To effectively handle the high-resolution input, we crop and weakly align each body part not only to handle pose variations but also to better recover fine details of the human body such as facial expressions. We demonstrated that the proposed method works effectively for the high-resolution images, for the various datasets.

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 (CVPR)*, 2019. 1

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus A. Magnor. Tex2shape: Detailed full human body geometry from a single image. *CoRR*, 2019. 3

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[4] AXYZ. https://secure.axyz-design.com/. 2

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE/CVF on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[7] Akin Caliskan, Armin Mustafa, and Adrian Hilton. Temporal consistency loss for high resolution textured and clothed 3d human reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[9] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 42:2523–2539, 2015. 1

[10] 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 (CVPR)*, 2021. 3

[11] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 1

[12] Epic Games. Unreal engine. 4

[13] Valentin Gabeur, Jean-Sébastien Franco Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 6

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[15] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[16] 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 (ICCV)*, 2021. 2

[17] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[18] 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 (CVPR)*, 2020. 2

[19] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 6

[20] Sai Sagar Jinka, Rohan Chacko, Avinash Sharma, and PJ Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 2

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[22] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 6

[24] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 3

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 34:1–16, 2015. 2, 3

[26] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, 1987. 2

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[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 on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[29] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[30] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[31] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018. 5

[32] Min-Gyu Park, Ju-Mi Kang, Je Woo Kim, and Ju Hong Yoon. Monocular human digitization via implicit re-projection networks. *arXiv preprint arXiv:2205.06468*, 2022. 5, 6

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE/CVF on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. 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 (CVPR)*, 2019. 8

[35] Marco Pesavento, Marco Volino, and Adrian Hilton. Super-resolution 3d human shape from a single low-resolution image. *arXiv preprint arXiv:2208.10738*, 2022. 1

[36] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3

[37] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

[38] Capturing Reality. Reality Capture. https://www.capturingreality.com/, 2022. 4

[39] Renderpeople. https://renderpeople.com/. 2, 6, 7

[40] 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 (ICCV)*, 2019. 1, 2, 7

[41] 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 (CVPR)*, 2020. 1, 3, 7

[42] Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. Self-supervised human depth estimation from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[43] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[44] Twindom. https://web.twindom.com/. 2

[45] Unity. UNLOCK YOUR CREATIVITY. https://unity.com/, 2022. 4

[46] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[47] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 4

[48] 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 (CVPR)*, 2022. 1, 3, 7

[49] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7

[50] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[51] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[52] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[53] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[54] 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 (TPAMI)*, 44:3170–3184, 2021. 1, 3, 7

[55] 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 (ICCV)*, 2019. 1, 2

[56] 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 (CVPR)*, 2019. 3

[57] Hao Zhu, Xinxin Zuo, Haotian Yang, Sen Wang, Xun Cao, and Ruigang Yang. Detailed avatar recovery from single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44:7363–7379, 2021. 3