Sampling is Matter: Point-guided 3D Human Mesh Reconstruction

Jeonghwan Kim[1*] Mi-Gyeong Gwon[1*] Hyunwoo Park[1]
Hyukmin Kwon[2] Gi-Mun Um[2] Wonjun Kim[1†]
[1]Konkuk University [2]Electronics and Telecommunications Research Institute
{jhkim0759,kmk3942,pzls,wonjkim}@konkuk.ac.kr {hmkwon,gmum}@etri.re.kr

Abstract

This paper presents a simple yet powerful method for 3D human mesh reconstruction from a single RGB image. Most recently, the non-local interactions of the whole mesh vertices have been effectively estimated in the transformer while the relationship between body parts also has begun to be handled via the graph model. Even though those approaches have shown the remarkable progress in 3D human mesh reconstruction, it is still difficult to directly infer the relationship between features, which are encoded from the 2D input image, and 3D coordinates of each vertex. To resolve this problem, we propose to design a simple feature sampling scheme. The key idea is to sample features in the embedded space by following the guide of points, which are estimated as projection results of 3D mesh vertices (i.e., ground truth). This helps the model to concentrate more on vertex-relevant features in the 2D space, thus leading to the reconstruction of the natural human pose. Furthermore, we apply progressive attention masking to precisely estimate local interactions between vertices even under severe occlusions. Experimental results on benchmark datasets show that the proposed method efficiently improves the performance of 3D human mesh reconstruction. The code and model are publicly available at: https://github.com/DCVL-3D/PointHMR_release.

1. Introduction

The goal of 3D human mesh reconstruction is to estimate 3D coordinates of points, which make up the human body surface. Since the high-quality 3D human model has been consistently required for various immersive applications, many studies have devoted considerable efforts to accurately reconstruct the 3D human mesh. In the early stage of this field, complex optimization techniques were adopted to generate the 3D human model based on the relationship between multiple scenes, which are acquired by using stereo or multiple-view camera systems. Recently, owing to the great success of deep learning, the problem of 3D human mesh reconstruction now can be resolved only with a single RGB image, thus the majority has begun to develop compact network architectures and efficient training strategies. Even though such deep learning-based approaches have shown the significant progress in 3D human mesh reconstruction, this task is still challenging due to severe occlusions by diverse human poses and depth ambiguities by the monocular setting.

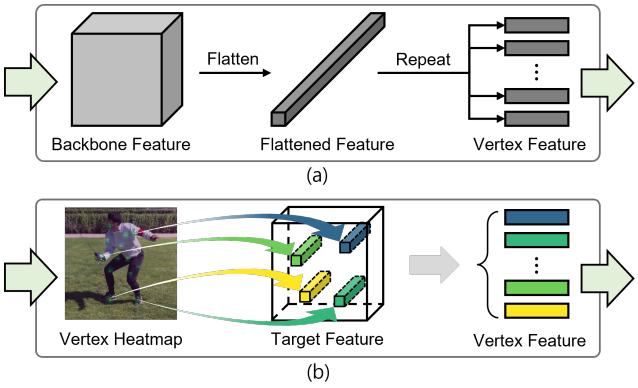

Figure 1. (a) Traditional process of feature extraction for estimating 3D coordinates. (b) Vertex-relevant feature extraction process based on the proposed point-guided sampling method for estimating 3D coordinates.

Deep learning-based approaches can be divided into two main groups: model-based and model-free methods. In the former, most methods aim to estimate shape and pose parameters of the skinned multi-person linear (SMPL) model [24], which is capable of yielding the whole vertices via these two simple factors, thus most widely employed in this field. Traditional encoder-decoder architectures, which are mostly composed of stacked convolutional layers, are sufficient to conduct the regression for estimating those parameters. Despite their great performance, model-based

*equal contribution
†corresponding author

methods have the obvious shortcoming, i.e., reconstruction results are limited to the pre-defined types of human body models. On the other hand, model-free methods have attempted to directly infer 3D coordinates of mesh vertices from input features without using any specific human body model. Compared to the model-based approach, which obtains the well-defined full mesh by adjusting shape and pose parameters, the model-free approach needs to estimate 3D coordinates of whole vertices directly from the network. Most methods in this category are based on the transformer to grasp non-local interactions between mesh vertices. The graph model (e.g., graph convolution) also has been utilized together to allow for body part relations in a local manner. One important advantage of the model-free approach is the flexibility to adapt to other applications, e.g., hand pose estimation, without significant changes of the data format and the training strategy. However, inferring the 3D coordinate from a single monocular image is still challenging due to lack of learning the correspondence between encoded features and spatial positions.

In this paper, we propose a simple yet powerful method for 3D human mesh reconstruction. To this end, we conduct feature sampling at vertex-relevant points of the input image as shown in Fig. 1, which are estimated through the heatmap decoder trained by projection results of 3D mesh vertices (i.e., ground truth). These sampled features are subsequently fed into the transformer encoder as the form of the vertex token (see Fig. 2). In a similar way of [6], we apply attention masking to the transformer encoder, however, the difference is that the local connection is defined with the range of multiple levels through the sequence of transformer encoders. This progressive attention masking helps the model understand local relations between vertices precisely even in occlusions. The main contribution of the proposed method can be summarized as follows:

- We propose to utilize the correspondence between encoded features and vertex positions, which are projected into the 2D space, via our point-guided feature sampling scheme. By explicitly indicating such vertex-relevant features to the transformer encoder, coordinates of the 3D human mesh are accurately estimated.

- Our progressive attention masking scheme helps the model efficiently deal with local vertex-to-vertex relations even under complicated poses and occlusions.

2. Related Work

In this Section, we give a brief review of the previous studies for 3D human mesh construction which have progressed in two different directions: model-based and model-free approaches.

Model-based approaches. As mentioned, most model-based approaches aim to estimate shape and pose parameters of the SMPL model [24] for restoring the entire set of mesh vertices. In the beginning, several studies attempted to align 2D joint positions as well as body part segments, which are estimated by respective networks, with the ground truth projected from the 3D human mesh [3, 19]. However, these methods require additional steps to estimate shape and pose parameters of the SMPL model. To cope with this limitation, Kanazawa *et al.* [12] proposed to regress such parameters directly from a single RGB image without using intermediate results, i.e., 2D joints and body part segments. Specifically, they designed a simple convolutional encoder with the adversarial loss to make the reconstructed mesh result be realistic. Inspired by the great potential of this simple regression scheme, many researchers have introduced various encoder-decoder architectures to estimate shape and pose parameters. Kolotouros *et al.* [17] proposed to combine the end-to-end regression model with the optimization loop to strongly supervise the refinement process. Choutas *et al.* [9] attempted to directly regress body, face, and hands by exploiting the body-driven attention in the SMPL-X [29] format for generating the high-quality 3D human mesh. Biggs *et al.* [2] designed a multi-hypothesis neural network regressor based on the best-of-M loss, which makes the plausible human pose even under severely occluded environments. Zhang *et al.* [37] also focused on accurately restoring object-occluded human shape and pose by utilizing the partial UV map and the novel saliency map in the SMPL format. Most recently, Kocabas *et al.* [16] devised the part attention module to guide the network to concentrate more on relevant body parts for inferring the given pose with SMPL parameters in a single input RGB image. Sun *et al.* [32] estimated body center positions instead of segmenting human regions and extracted the parameter maps for the SMPL model at the corresponding positions. They further extend their algorithm by adopting the heat map of the bird-eye view to alleviate the depth ambiguity in monocular settings [33]. On the other hand, the kinematic topology module has been embedded into the neural network architecture to consider the relationship between articulations of the human body [20]. Even though model-based approaches have shown the remarkable progress in 3D human mesh reconstruction, their performance is limited to pre-defined types of human body models, which are hardly extended to other applications.

Model-free approaches. Model-free approaches intend to directly restore the entire set of mesh vertices while relaxing the heavy reliance on the parameter space in model-based methods. As a pioneer, Kolotouros *et al.* [18] proposed to estimate human mesh coordinates through the graph convolutional neural network (GraphCNN). To do this, they attached encoded features to nodes in the graph, which are

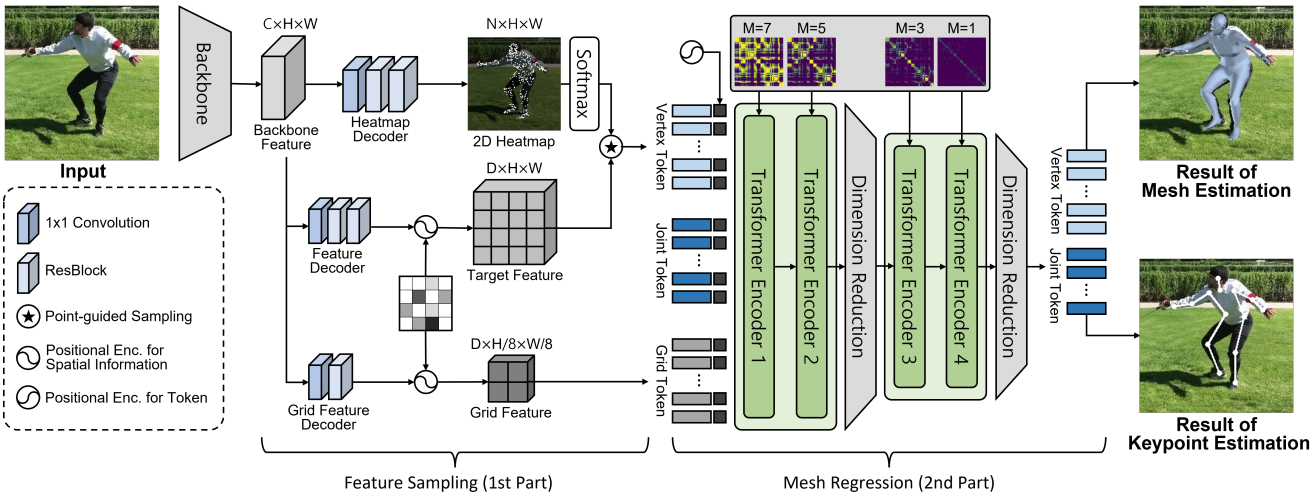

Figure 2. Overall architecture of the proposed method. The backbone feature is used to generate heatmap, target feature, and grid feature with respective decoders. Our sampling scheme utilizes heatmap and target feature to make the vertex token. Note that N and M denote the number of vertices and the distance threshold for defining the local connection in self-attention, respectively.

mapped to 3D coordinates of the template mesh. From this perspective, Choi *et al.* [7] also adopted GraphCNN to reconstruct 3D human meshes from 2D and 3D pose information in a coarse-to-fine manner. While these GraphCNN-based methods have a good ability to fully exploit the mesh topology, they are somewhat lacking in considering global interactions between joints and vertices. To cope with this limitation, the transformer has begun to be actively adopted for model-free approaches. Specifically, Lin *et al.* [21] firstly introduced the transformer encoder, which simply takes joint and vertex queries as input tokens, to regress 3D coordinates from a single input RGB image. In particular, they further embedded GraphCNN into the transformer block to supplement local interactions, e.g., between-part relationships, for 3D human mesh reconstruction. Most recently, Cho *et al.* [6] have disentangled image features and mesh queries by utilizing the transformer encoder-decoder architecture to alleviate the high complexity of interactions among input tokens.

Our method is also based on the transformer encoder with a simple sampling scheme, which gives a great help to focus on vertex-relevant features for inferring coordinates of the 3D human mesh. Technical details will be explained in the following Section.

3. Proposed Method

The proposed method consists of two main parts. Specifically, vertex-relevant features are extracted based on our point-guided feature sampling in the first part while 3D coordinates are estimated through the sequence of transformer encoders with the proposed progressive attention masking scheme in the second part. The overall architecture of the proposed method is illustrated in Fig. 2.

Figure 3. Several examples of predicted heatmaps on the 3DPW dataset. For better visibility, the activated regions of the heatmap corresponding to several selected vertices are represented in a single image with bright colors (best view in colors).

3.1. Point-guided Feature Sampling

Since the direct transform from the color value to the 3D coordinate is still a difficult process due to heterogeneous modalities, we propose to use the intermediate guidance, i.e., features sampled at positions of vertices projected from 3D to 2D spaces. Specifically, we represent such projection results as the form of the heatmap, and sample features at activated positions in this heatmap. The detailed process of our point-guided feature sampling is shown in the first part of Fig. 2. Firstly, the backbone feature $X_b \in \mathbb{R}^{C \times H \times W}$ is encoded through the backbone network (HRNet [35] in this work), where C, H, and W denote the number of channels, height, and width, respectively.

X_b is subsequently decoded into the heatmap, the target feature, and the grid feature by respective decoders as shown in Fig. 2. After that, the vertex-relevant feature is sampled at each of N vertices based on combination of the predicted heatmap and the target feature. Several examples of the predicted heatmap are shown in Fig. 3. Note that we conducted element-wise multiplication of the predicted heatmap and the target feature to make our sampling process be differentiable. Moreover, we apply the positional

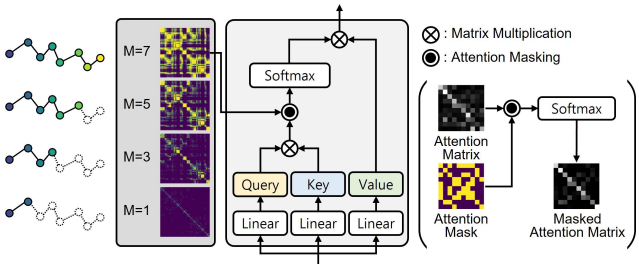

Figure 4. The detailed structure of the proposed progressive attention masking scheme. Note that this figure illustrates the masking process in the first transformer encoder, i.e., the case of $M = 7$, shown in Fig. 2.

encoding to the target feature in the same way of [5, 6] for preserving the spatial information in the transformer encoder. Our point-guided feature sampling can be formulated as follows:

$$\hat{V}_i = H_i \times F, \quad i = 1, 2, 3, ..., N, \tag{1}$$

where H_i and F denote the predicted heatmap corresponding to the i-th vertex and the target feature, respectively. N denotes the total number of vertices and is set to 431 as used in previous methods. $\hat{V}_i$ is the sampling result, which is finally flattened and fed into the transformer encoder with the grid feature in the second part of our proposed method. It is noteworthy that the grid feature plays an important role to create the united body structure by aligning each point in an appropriate location.

3.2. Transformer Encoder with Progressive Attention Masking

In this subsection, we explain the sequence of transformer encoders in detail. In a similar way of [21, 22], we design the transformer encoder with the dimension reduction layer as shown in the second part of Fig. 2. Specifically, the transformer encoder takes the vertex token $\hat{V} \in \mathbb{R}^{N \times D}$, the joint token $\hat{J} \in \mathbb{R}^{K \times D}$, and the grid token $\hat{G} \in \mathbb{R}^{Z \times D}$ as inputs. The joint token is a trainable parameter that is randomly initialized and optimized during the training phase whereas vertex and grid tokens are extracted from the first part of the proposed method. Moreover, positional encodings are employed to give the identity by concatenating the trainable parameter to each token. The dimension of the encoded token is reduced by linear projection after each transformer block, which consists of two transformer encoders. To consider local vertex-vertex relations as well as non-adjacent interactions, we exclude far-distant connections between vertices to compute self-attention in the transformer encoder. In contrast to the previous method [6], we gradually decrease the range to define the local connection between vertices through the sequence of transformer encoders as illustrated in Fig. 4. This helps the model consider the local relationship between neighbor vertices in a progressive manner. Figure 5 shows the effect of our progressive attention masking. As can be seen, the proposed method provides the reasonable result even in occlusions.

(a) (b) (c)

Figure 5. (a) Input image. (b) Reconstruction result with previous attention masking [6]. (c) Reconstruction result with progressive attention masking (proposed).

The outputs (i.e., vertex and joint tokens) of the last transformer encoder are finally projected into 3D coordinates via a linear layer. In the case of the vertex token, the upsampling algorithm introduced in [31] is applied to expand sparse vertices $V' \in \mathbb{R}^{N \times 3}$ into dense vertices $V \in \mathbb{R}^{\mathcal{N} \times 3}$ as follows:

$$V = UV', \tag{2}$$

where U is the pre-defined upsampling matrix [31]. $\mathcal{N}$ is set to 6,890 (same as the vertex number of the SMPL model). The predicted 3D human mesh and 3D keypoints can be visualized as the rightmost images shown in Fig. 2.

3.3. Loss Function

The proposed method is trained based on four types of loss functions, i.e., vertex loss $\mathcal{L}_v$, 3D joint loss $\mathcal{L}_{j3d}$, 2D joint loss $\mathcal{L}_{j2d}$, and heatmap loss $\mathcal{L}_h$. The first three loss functions are used for estimating positions of vertices and joints as introduced in previous works [6, 21, 22] while the last one guides the network to find positions of projected vertices. First of all, L_1 loss is adopted to compute the difference between positions of the predicted vertex V and the corresponding ground truth $\tilde{V}$ as follows:

$$\mathcal{L}_v = \frac{1}{N} \sum_{i=1}^{N} \| \tilde{V}^i - V^i \|_1, \tag{3}$$

where N denotes the total number of vertices. For the 3D joint loss, we compute the distance between the estimated joint position J_{3d} and the corresponding ground truth $\tilde{J}_{3d}$ in the 3D space. Note that the joint position, which is regressed from the predicted vertices, i.e., $\bar{J}_{3d}$, is also used for the loss computation as follows:

$$\mathcal{L}_{j3d} = \frac{1}{K} \sum_{i=1}^{K} \| \tilde{J}^i_{3d} - J^i_{3d} \|_2 + \| \tilde{J}^i_{3d} - \bar{J}^i_{3d} \|_2, \tag{4}$$

where K denotes the total number of joints. Similarly, the 2D joint loss is calculated as well. To do this, J_{3d} and $\bar{J}_{3d}$

	Methods	Backbone	Human3.6M		3DPW		
			MPJPE(↓)	PA-MPJPE(↓)	MPJPE(↓)	PA-MPJPE(↓)	MPVPE (↓)
model-based	HMR [12]	ResNet-50	88.0	56.8	130.0	81.3	–
	SPIN [17]	ResNet-50	–	41.1	96.9	59.2	116.4
	ExPose [9]	ResNet-50	–	–	93.4	55.6	–
	VIBE [15]	ResNet-50	65.6	41.4	82.9	51.9	99.1
	HybrIK [20]	ResNet-34	54.4	34.5	80.0	48.8	94.5
	ROMP [32]	HRNet-W32	–	–	76.7	47.3	93.4
	PARE [16]	HRNet-W32	–	–	74.5	46.5	88.6
	MAED [40]	ResNet-50	56.4	38.7	79.1	45.7	92.6
	PyMAF [36]	ResNet-50	57.7	40.5	92.8	58.9	110.1
	BEV* [33]	HRNet-W32	–	–	78.5	46.9	92.3
	OCHMR* [13]	ResNet-50 / HRNet-W32	–	–	89.7	58.3	107.1
	3DCrowdNet* [8]	ResNet-50	–	–	81.7	51.5	98.3
model-free	GraphCMR [18]	ResNet-50	–	50.1	–	70.2	–
	I2LMeshNet [27]	ResNet-50	55.7	41.7	93.2	57.7	110.1
	Pose2Mesh [7]	HRNet-W48	64.9	47.0	89.5	56.3	105.3
	METRO [21]	HRNet-W64	54.0	36.7	77.1	47.9	88.2
	MeshGraphormer [22]	HRNet-W64	51.2	34.5	74.7	45.6	87.7
	FastMETRO [6]	HRNet-W64	52.2	33.7	**73.5**	**44.6**	**84.1**
	Ours	HRNet-W32	**48.3**	**32.9**	73.9	44.9	85.5

Table 1. Performance comparisons of 3D human mesh reconstruction based on Human3.6M and 3DPW datasets. The proposed method achieves the best performance in the Human3.6M dataset while still showing the competitive performance in the 3DPW dataset (best results are shown in bold). Note that * denotes the performance without fine-tuning on the 3DPW dataset.

are projected onto the 2D space and the corresponding results are represented as J_{2d} and $\bar{J}_{2d}$, respectively. Based on such projected results, the 2D joint loss is formulated as L_2 loss in the same way of the 3D joint loss as follows:

$$\mathcal{L}_{j2d} = \frac{1}{K}\sum_{i=1}^{K} \|\tilde{J}^i_{2d} - J^i_{2d}\|_2 + \|\tilde{J}^i_{2d} - \bar{J}^i_{2d}\|_2, \quad (5)$$

where $\tilde{J}_{2d}$ is the ground truth of the 2D joint position. On the other hand, our heatmap loss $\mathcal{L}_h$ consists of the binary cross entropy loss and the dice loss. Specifically, we adopt the binary cross entropy loss to determine whether the activated position is matched to the point projected from the vertex or not as follows:

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{i=1}^{N} \tilde{H}^i \log H^i + (1-\tilde{H}^i)\log(1-H^i), \quad (6)$$

where H and $\tilde{H}$ denote the predicted heatmap and the corresponding ground truth, respectively. The ground truth is represented as the binary map where the position of the projected vertex is assigned 1, otherwise 0. To alleviate the data-imbalanced problem, i.e., the projected point exists on only a single pixel in the ground truth image, the dice loss [26] is also employed as follows:

$$\mathcal{L}_{dice} = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{2\times(\tilde{H}^i \times H^i)}{\tilde{H}^i + H^i}. \quad (7)$$

Since the dice loss mainly focuses on the overlapped area, the data-imbalanced problem can be efficiently alleviated. By using the combination of these loss functions, the proposed network successfully learns to reconstruct the 3D human mesh as follows:

$$\begin{aligned}\mathcal{L}_{total} = {} & w_v\mathcal{L}_v + w_{j3d}\mathcal{L}_{j3d} + w_{j2d}\mathcal{L}_{j2d} \\ & + w_{bce}\mathcal{L}_{bce} + w_{dice}\mathcal{L}_{dice},\end{aligned} \quad (8)$$

where w_v, w_{j3d}, w_{j2d}, w_{bce} and w_{dice} are the balancing factor for each loss term, which are set to 0.01, 0.1, 0.01, 1.0, and 0.001, respectively.

4. Experimental Results

4.1. Training

All the experiments are implemented on the PyTorch framework [28] with an Intel E5-1650@3.60GHz CPU and two NVIDIA RTX A6000 GPUs. To train all the parameters of our model, the Adam optimizer [14] is adopted where the momentum factors are set to 0.9 and 0.999, respectively. The proposed network is trained for 50 epochs with a batch size of 64 per GPU. The learning rate is firstly set to 1×10^{-4} and reduced to 1×10^{-5} at the half of the learning time. For each input image, the area including the human is cropped and resized to the resolution of 224×224 pixels before training.

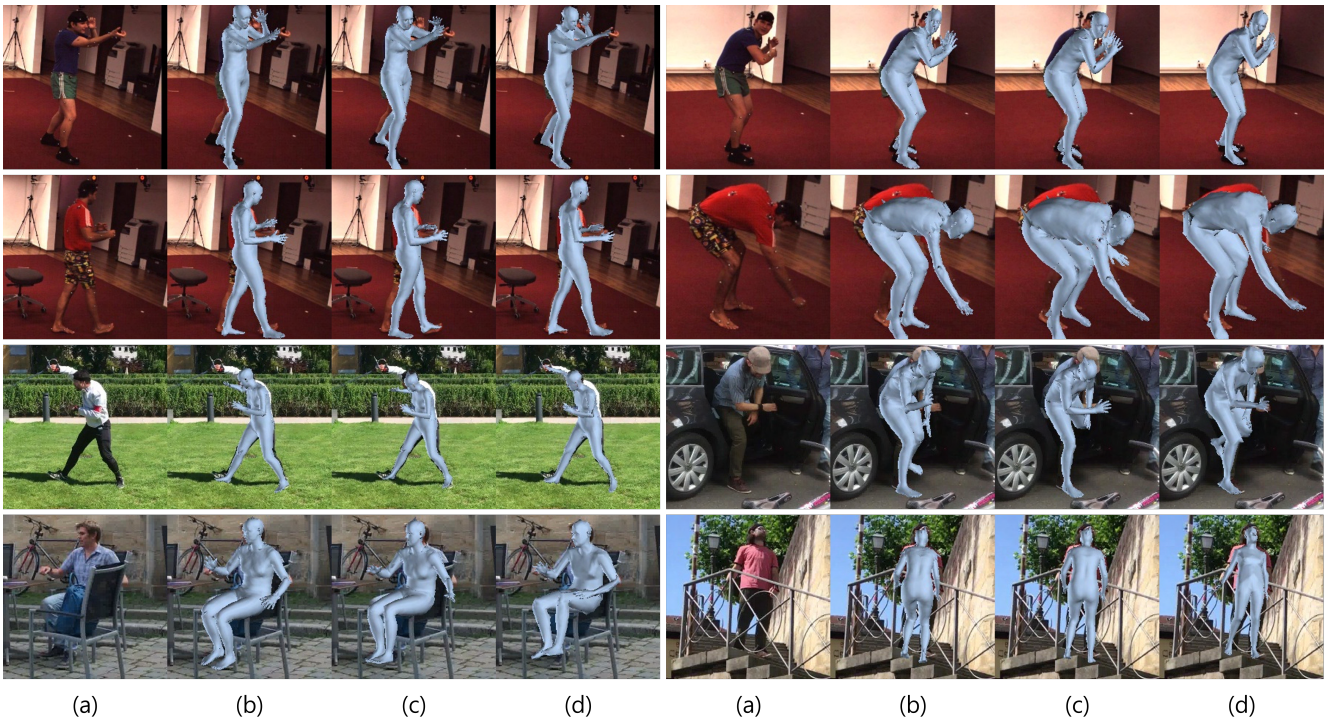

Figure 6. Results of 3D human mesh reconstruction on Human3.6M [11] (top-two rows) and 3DPW [34] (bottom-two rows) datasets. (a) Input images. (b) Results by METRO [21]. (c) Results by Mesh Graphormer [22]. (d) Results by the proposed method.

4.2. Datasets and Evaluation Metrics

Datasets. For the performance evaluation of the proposed method, two representative benchmarks, i.e., Human3.6M [11] and 3DPW [34], are employed. Specifically, the proposed method is trained based on five datasets, i.e., Human3.6M [11], MuCo-3DHP [25], UP-3D [19], COCO [23], and MPII [1], by following previous approaches, and tested with the P2 protocol in the Human3.6M dataset. Since the ground truth of the 3D human mesh is unavailable in the Human3.6M dataset, we use the pseudo mesh label generated by SMPLify-X [29] as introduced in [6, 7, 21, 22, 27]. For the test on the 3DPW dataset, we fine-tune the proposed method by using the training set of the 3DPW dataset.

Evaluation metrics. For the quantitative evaluation, we use three metrics, i.e., mean per joint position error (MPJPE) [11], Procrustes-aligned mean per joint position error (PA-MPJPE) [38], and mean per vertex position error (MPVPE) [30], which have been widely adopted for the performance comparison in this field. Specifically, MPJPE measures the average value of the Euclidean distance between each estimated 3D joint and the corresponding ground truth. PA-MPJPE indicates MPJPE in which the estimated human body is aligned in terms of rotation and scaling using the Procrustes analysis. On the other hand, MPVPE is a metric for computing the Euclidean distance between coordinates of the predicted vertex and the corresponding ground truth.

4.3. Performance Evaluation

Quantitative evaluation. To demonstrate the efficiency and robustness of the proposed method, we compare ours with representative methods for 3D human mesh reconstruction, i.e., HMR [12], SPIN [17], ExPose [9], VIBE [15], HybrIK [20], ROMP [32], PARE [16], MAED [40], PyMAF [36], BEV [33], OCHMR [13], 3DCrowdNet [8], GraphCMR [18], I2LMeshNet [27], Pose2Mesh [7], METRO [21], Mesh Graphormer [22], and FastMETRO [6]. The result of the performance comparison is shown in Table 1. As can be seen, the proposed method achieves 48.3 MPJPE and 32.9 PA-MPJPE on the Human3.6M dataset, which outperforms the state-of-the-art methods with the meaningful performance gain. Even though the performance of the proposed method is slightly dropped compared to the best one (i.e., FastMETRO) in the 3DPW dataset, our method still shows the competitive performance with state-of-the-art methods. Specifically, model-free methods have shown the reliable performance without using the well-defined human model in recent days. This is because their reconstruction results are not limited to the small set of pre-defined human models, thus show more appropriate human meshes for a given image compared to model-based methods. In particular, transformer-based ar-

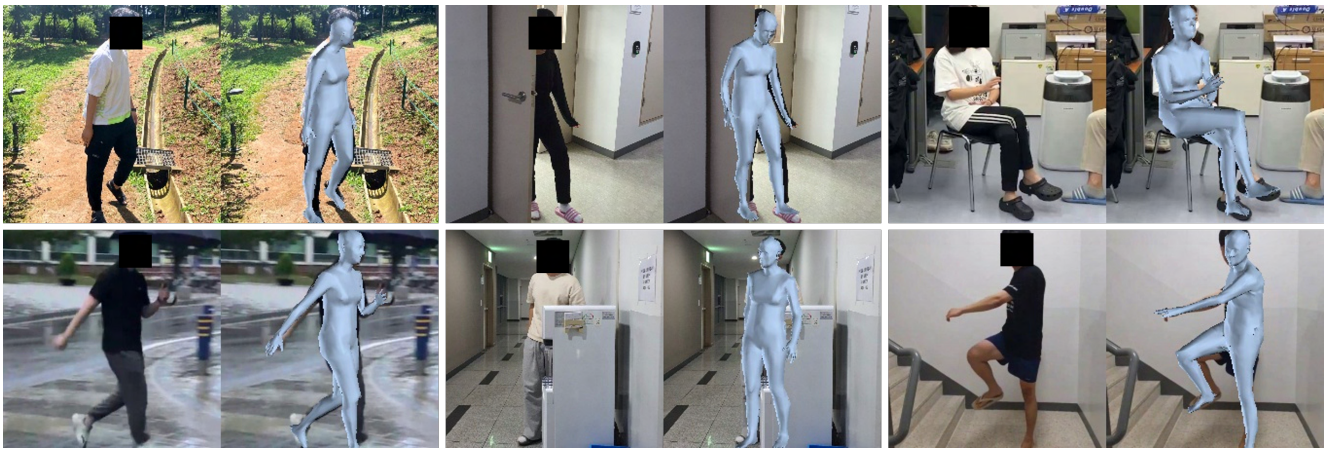

Figure 7. More results of 3D human mesh reconstruction for our samples, which are acquired by the smartphone. Odd columns: input images. Even columns: our results. Note that the proposed method provides reliable results of 3D human mesh reconstruction for various pictures of our daily life even with severe occlusions.

chitectures, e.g., Mesh Graphormer and FastMETRO, significantly improves the performance of 3D human mesh reconstruction. Nevertheless, recent model-based methods still show the competitive performance and work robust to occlusion cases, e.g., BEV, OCHMR, and 3DCrowdNet. It is noteworthy that the proposed method notably improves the performance by our point-guided feature sampling scheme without specially designing the decoder architecture like [6]. Moreover, the proposed method provides the reliable performance with the relatively small-sized backbone (e.g., HRNet-W32).

Qualitative evaluation. Several results of 3D human mesh reconstructions for Human3.6M and 3DPW datasets are shown in Fig. 6. Note that our results are reconstructed based on the upsampling matrix as used in [6] while other two methods, i.e., METRO [21] and Mesh Graphormer [22], utilized two linear layers to upsample coarse vertex points. We can see that the proposed method successfully estimates 3D human poses under various situations including real-world scenarios in outdoor scenes as well as the controlled environment. Specifically, previous methods have somewhat difficulties to estimate unusual poses, e.g., overlapping arms and bending, whereas the proposed method provides the well-fit 3D model for a given image. Moreover, the proposed method has a good ability to reconstruct 3D human meshes in occlusions due to the progressive attention masking scheme as shown in examples of the last row of Fig. 6. In particular, previous methods fail to infer the global orientation of the human body due to the fence in front of the target person, which leads to the significant performance drop for 3D human mesh reconstruction. In contrast, the proposed method shows the reliable performance with various occlusions (see third and fourth rows in Fig. 6). More examples for 3D human mesh reconstruction by the proposed method are shown in Fig. 7. Note that input images are acquired by the smartphone. As can be seen, the proposed method performs well for various pictures of our daily life even with severe occlusions. Therefore, it is thought that the proposed method paves the way for 3D human mesh reconstruction under various environments.

Point-guided feature sampling	Progressive attention masking	MPJPE	PA-MPJPE
✗	✗	63.2	39.9
✗	✓	61.2	40.8
✓	✗	50.9	33.3
✓	✓	**48.3**	**32.9**

Table 2. Performance analysis of the proposed method according to changes in the network architecture based on the Human3.6M dataset (best results are shown in bold).

Methods	MPJPE	PA-MPJPE
Without attention masking	50.9	33.3
Single attention masking	49.1	33.5
Progressive attention masking (ours)	**48.3**	**32.9**

Table 3. Performance analysis of the proposed method according to the attention masking scheme based on the Human3.6M dataset (best results are shown in bold). Note that $M = 1$ is used for single attention masking.

4.4. Ablation Studies

In this subsection, we first demonstrate the comparative experimental results by changing the components of the proposed method based on the Human3.6M dataset. Table 2 shows the contribution of such components. As can be seen, the performance of 3D human mesh reconstruc-

Figure 8. Several reconstruction results (3D joints (left) and 3D hand meshes (right)) by the proposed method for the FreiHAND [39] dataset. Note that the proposed method works robust to self-occlusions frequently occurring by complicated hand poses.

tion is significantly improved with our point-guided feature sampling scheme (MPJPE: 63.2 $\rightarrow$ 50.9, PA-MPJPE: 39.9 $\rightarrow$ 33.3). From this analysis, we can see that feature sampling at vertex-relevant positions even in the 2D space is effective for improving the performance of 3D human mesh reconstruction. In what follows, we also check the effect of our progressive attention masking and the corresponding result is shown in Table 3. Note that our point-guided feature sampling is applied for this experiment. By comparing ours with baseline (i.e., without masking) and single masking attention scheme [6], we can see that the progressive restriction strategy in defining the local connection is also helpful for improving the performance of 3D human mesh reconstruction. This tells us that the combination of our contributions makes the model be robust to complicated real-world environments.

4.5. Generalization Ability

In contrast to the model-based approach, the proposed method can be easily applied to other applications. To show the generalization ability of the proposed method, we conduct 3D hand mesh reconstruction based on the FreiHAND [39] dataset by changing the number of input tokens for the sequence of transformer encoders. Note that the number of vertices to be projected is set to 195 and those are upsampled to 778 via the same upsampling matrix used for 3D human mesh reconstruction. Several reconstruction results by the proposed method are shown in Fig. 8. We can see that the proposed method provides the reliable reconstruction results (i.e., 3D joints and 3D hand meshes) for various hand poses. In particular, self-occlusions by complicated relations between adjacent fingers are successfully handled in the proposed method. The result of the quantitative evaluation is also shown in Table 4. As can be seen, the proposed method shows the competitive performance on the FreiHAND dataset.

Methods	PA-MPVPE	PA-MPJPE	F@5mm	F@15mm
Hasson *et al.* [10]	13.2	–	0.436	0.908
Boukhayma *et al.* [4]	13.0	–	0.435	0.898
FreiHAND [39]	10.7	–	0.529	0.935
I2LMeshNet [27]	7.6	7.4	0.681	0.973
Pose2Mesh [7]	7.8	7.7	0.674	0.969
METRO [21]	6.7	6.8	0.717	0.981
FastMETRO [6]	–	6.5	–	0.982
Ours	**6.6**	**6.1**	**0.720**	**0.984**

Table 4. Performance comparisons of 3D hand mesh reconstruction based on the FreiHAND dataset (best results are shown in bold).

5. Conclusion

In this paper, we present a simple yet powerful method for 3D human mesh reconstruction from a single RGB image. The key idea of the proposed method is to alleviate heterogeneous modalities between input (i.e., color) and output (i.e., coordinate) by considering the correspondence of encoded features and 2D points projected from 3D vertices. Our point-guided feature sampling scheme is to sample vertex-relevant features based on the combination of the heatmap and encoded features. In addition, the proposed progressive attention masking scheme makes the model to be robust to occlusions by considering local connections of different levels through the sequence of transformer encoders. Experimental results on benchmark datasets show that the proposed method performs reliably for various real-world environments.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (2021-0-02084, eXtended Reality and Volumetric media generation and transmission technology for immersive experience sharing in noncontact environment with a Korea-EU international cooperative research).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3686–3693, 2014. 6

[2] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3D Multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. In *Adv. Neural Inform. Process. Syst.*, pages 20496–20507, 2020. 2

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Eur. Conf. Comput. Vis.*, pages 561–578, 2016. 2

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10843–10852, 2019. 8

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229, 2020. 4

[6] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Eur. Conf. Comput. Vis.*, pages 342–359, 2022. 2, 3, 4, 5, 6, 7, 8

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *Eur. Conf. Comput. Vis.*, pages 769–787, 2020. 3, 5, 6, 8

[8] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1475–1484, 2022. 5, 6

[9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *Eur. Conf. Comput. Vis.*, pages 20–40, 2020. 2, 5, 6

[10] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11807–11816, 2019. 8

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 6

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018. 2, 5, 6

[13] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1715–1725, 2022. 5, 6

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, pages 1–13, 2015. 5

[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5253–5263, 2020. 5, 6

[16] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Int. Conf. Comput. Vis.*, pages 11127–11137, 2021. 2, 5, 6

[17] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Int. Conf. Comput. Vis.*, pages 2252–2261, 2019. 2, 5, 6

[18] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4501–4510, 2019. 2, 5, 6

[19] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6050–6059, 2017. 2, 6

[20] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3383–3393, 2021. 2, 5, 6

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1954–1963, 2021. 3, 4, 5, 6, 7, 8

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, pages 12939–12948, 2021. 4, 5, 6, 7

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 6

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1, 2

[25] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Int. Conf. 3D Vis.*, pages 120–130, 2018. 6

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. Int. Conf. 3D Vis.*, pages 565–571, 2016. 5

[27] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *Eur. Conf. Comput. Vis.*, pages 752–768, 2020. 5, 6, 8

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Adv. Neural Inform. Process. Syst.*, pages 1–4, 2017. 5

[29] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10975–10985, 2019. 2, 6

[30] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 459–468, 2018. 6

[31] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *Eur. Conf. Comput. Vis.*, pages 704–720, 2018. 4

[32] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 2, 5, 6

[33] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13243–13252, 2022. 2, 5, 6

[34] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Eur. Conf. Comput. Vis.*, pages 601–617, 2018. 6

[35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2021. 3

[36] 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 *Int. Conf. Comput. Vis.*, 2021. 5, 6

[37] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7376–7385, 2020. 2

[38] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):901–914, 2018. 6

[39] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 813–822, 2019. 8

[40] Maoqing Tian Jianbo Liu Shuai Yi Hongsheng Li Ziniu Wan, Zhengjia Li. Encoder-decoder with multi-level attention for 3D human shape and pose estimation. In *Int. Conf. Comput. Vis.*, 2021. 5, 6