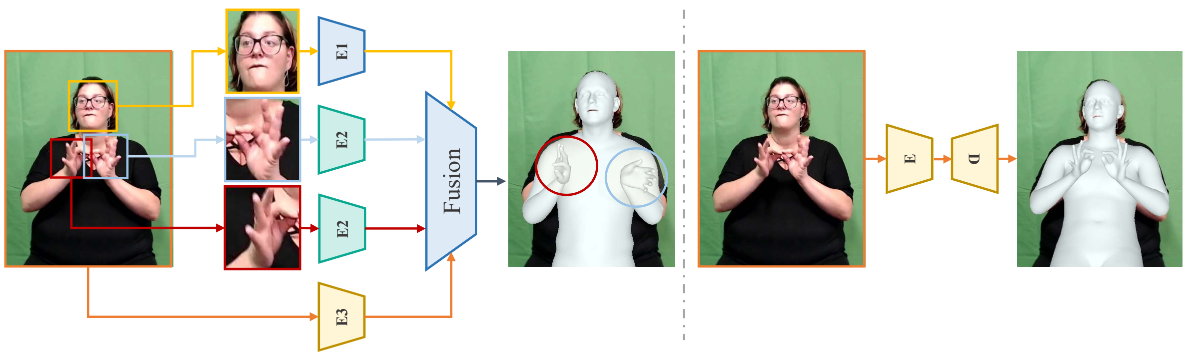
One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer

Jing Lin[1,2§], Ailing Zeng[1¶], Haoqian Wang[2], Lei Zhang[1], Yu Li[1]
[1] International Digital Economy Academy (IDEA),
[2] Shenzhen International Graduate School, Tsinghua University
https://osx-ubody.github.io



(a) Previous multi-stage pipeline (b) Our one-stage pipeline

Figure 1. A comparison of existing whole-body mesh recovery methods and ours. Most existing methods leverage a multi-stage pipeline which uses separate expert models to process body component (*e.g.*, **E1**: HeadNet, **E2**: HandNet, **E3**: BodyNet) and fuse them to get the whole-body prediction in a copy-paste manner. The result (from [36]) produces unnatural wrist poses. In contrast, our pipeline is a neat one-stage framework with a single encoder-decoder and can predict more accurately with natural meshes.

Abstract

Whole-body mesh recovery aims to estimate the 3D human body, face, and hands parameters from a single image. It is challenging to perform this task with a single network due to resolution issues, i.e., the face and hands are usually located in extremely small regions. Existing works usually detect hands and faces, enlarge their resolution to feed in a specific network to predict the parameter, and finally fuse the results. While this copy-paste pipeline can capture the fine-grained details of the face and hands, the connections between different parts cannot be easily recovered in late fusion, leading to implausible 3D rotation and unnatural pose. In this work, we propose a one-stage pipeline for expressive whole-body mesh recovery, named OSX, without separate networks for each part. Specifically, we design a Component Aware Transformer (CAT) composed of a global body encoder and a local face/hand decoder. The encoder predicts the body parameters and provides a high-quality feature map for the decoder, which performs a feature-level upsample-crop scheme to extract high-resolution part-specific features and adopt keypoint-guided deformable attention to estimate hand and face precisely. The whole pipeline is simple yet effective without any manual post-processing and naturally avoids implausible prediction. Comprehensive experiments demonstrate the effectiveness of OSX. Lastly, we build a large-scale Upper-Body dataset (UBody) with high-quality 2D and 3D whole-body annotations. It contains persons with partially visible bodies in diverse real-life scenarios to bridge the gap between the basic task and downstream applications.

1. Introduction

Expressive whole-body mesh recovery aims to jointly estimate the 3D human body poses, hand gestures, and facial expressions from monocular images. It is gaining increasing attention due to recent advancements in whole-body parametric models (*e.g.*, SMPL-X [37]). This task is a key step in modeling human behaviors and has many applications, *e.g.*, motion capture, human-computer interaction. Previous research focus on individual tasks of reconstructing human body [9, 21, 25, 44, 48, 49], face [2, 10, 12, 43],

§ Work done during an internship at IDEA; ¶ Corresponding author.

or hand [4, 8, 16]. However, whole body mesh recovery is particularly challenging as it requires accurate estimation of each part and natural connections between them.

Existing learning-based works [13, 29, 36, 41, 51] use multi-stage pipelines for body, hand, and face estimation to achieve the goal of this task. As depicted in Figure 1(a), these methods typically detect different body parts, crop and resize each region, and feed them into separate expert models to estimate the parameters of each part. The multi-stage pipeline with different estimators for body, hand, and face results in a complicated system with a large computational complexity. Moreover, the blocked communications among different components inevitably cause incompatible configurations, unnatural articulation of the mesh, and implausible 3D wrist rotations as they cannot obtain informative and consistent clues from other components. Some methods [13,29,51] attempt to alleviate these issues by designing additional complicated integration schemes or elbow-twist compensation fusion among individual body parts. However, these approaches can be regarded as a late fusion strategy and thus have limited ability to enhance each other and correct implausible predictions.

In this work, we propose a one-stage framework named *OSX* for 3D whole-body mesh recovery, as shown in Figure 1(b), which does not require separate networks for each part. Inspired by recent advancements in Vision Transformers [11, 47], which are effective in capturing spatial information in a plain architecture, we design our pipeline as a component-aware Transformer (CAT) composed of a global body encoder and a local component-specific decoder. The encoder equipped with body tokens as inputs captures the global correlation, predicts the body parameters, and simultaneously provides high-quality feature map for the decoder. The decoder utilizes a differentiable upsample-crop scheme to extract part-specific high-resolution features and adopt the keypoint-guided deformable attention to precisely locate and estimate hand and face parameters. The proposed pipeline is simple yet effective without any manual post-processing. To the best of our knowledge, this is the first one-stage pipeline for 3D whole-body estimation. We conduct comprehensive experiments to investigate the effects of the above designs and compare our method, with existing works on three benchmarks. Results show that *OSX* outperforms the state-of-the-art (SOTA) [29] by 9.5% on AGORA, 7.8% on EHF, and 13.4% on the body-only 3DPW dataset.

In addition, existing popular benchmarks, as illustrated in the first row of Figure 2, are either indoor single-person scenes with limited images (*e.g.*, EHF [37]) or outdoor synthetic scenes (e.g., AGORA [35]), where the people are often too far from the camera and the hands and faces are frequently obscured. In fact, human pose estimation and mesh recovery is a fundamental task that benefits many downstream applications, such as sign language recognition, gesture generation, and human-computer interaction. Many scenarios, such as talk shows and online classes, are of vital importance to our daily life yet under-explored. In such scenarios, the upper body is a major focus, whereas the hand and face are essential for analysis. To address this issue, we build a large-scale upper-body dataset with fifteen human-centric real-life scenes, as shown in Figure 2(f) to (t). This dataset contains many unseen poses, diverse appearances, heavy truncation, interaction, and abrupt shot changes, which are quite different from previous datasets. Accordingly, we design a systematical annotation pipeline and provide precise 2D whole-body keypoint and 3D whole-body mesh annotations. With this dataset, we perform a comprehensive benchmarking of existing whole-body estimators.

Our contributions can be summarized as follows.

- We propose a one-stage pipeline, *OSX*, for 3D whole-body mesh recovery, which can regress the SMPL-X parameters in a simple yet effective manner.
- Despite the conceptual simplicity of our one-stage framework, it achieves the new state of the art on three popular benchmarks.
- We build a large-scale upper-body dataset, *UBody*, to bridge the gap between the basic task and downstream applications and provide precise annotations, with which we conduct benchmarking of existing methods. We hope *UBody* can inspire new research topics.

2. Related Work

2.1. Methods of Whole-body Mesh Recovery

Whole-body mesh recovery targets to localize mesh vertices of all human components, including body, hands, and face from monocular images. Most previous works focus only on individual hand [4, 8, 16], face [2, 10, 12, 43], and body [23–25, 44, 48] reconstruction. In contrast, the joint whole-body estimation methods are less addressed. Some optimization-based works reconstruct 3D bodies by fitting the detected 2D keypoints from images with additional constraints, but they are slow and prone to local optima [37,46]. Thanks to the whole-body parametric model (*e.g.*, SMPL-X [37]), learning-based models [13, 36, 41, 42, 54] emerge to train networks to predict expressive body pose, shape, hand gesture, and facial expression. Due to the low resolution of hands and face, these whole-body methods crop and resize the hands and face images to higher resolutions and feed them into separate expert networks to conduct the corresponding parameter regression. Specifically, ExPose [36] introduces body-driven attention for higher-resolution crops of the face and hand estimation, a dedicated refinement module, and part-specific knowledge from existing hand-only and face-only datasets. FrankMocap [41] presents a regression-and-integration method to build a fast and accurate system. PIXIE [13] produces animatable whole body

Figure 2. Illustration of five previous datasets (from (a) to (e)) and the proposed Upper Body Dataset (from (f) to (t)) with fifteen real-life scenes. *UBody* bridges the gap between the basic 3D whole-body estimation task and downstream tasks with highly expressive actions.

with realistic facial details via a moderator to fuse body part features adaptively. Recently, Hand4Whole [29] utilizes both body and hand joint features for accurate 3D wrist rotation and smooth connection between body and hands.

Nevertheless, these methods aim at high performance by using separate networks in a divide-and-conquer fashion for different components and a specific fusion module to paste them together. The multi-stage pipelines lead to high complexity and inevitably cause inconsistent and unnatural articulation of the mesh and implausible 3D wrist rotations, especially in occluded, truncated, and blurry contexts. Until now, one-stage methods in this task are unexplored.

2.2. Benchmarks of Expressive Body

Some datasets with parametric model annotations [5, 17, 20, 30, 35, 37, 45] have been developed to advance the field. Table 1 summarizes these datasets from the annotation type, size, scene diversity, *etc*. To be specific, EHF [37] is the first evaluation dataset for SMPL-X-based models, which is built by capturing 3D body shapes with a scanning system and then fitting the SMPL-X model to the scans. AGORA [35] is a synthetic dataset with high realism and accurate ground truth, which is by far the most commonly used test data due to the diversity of subjects, environments, clothes, and occlusions. Notably, people in AGORA are often far from the camera, and their hands and face are obscured and have small resolutions, making existing methods focus more on body rather than hand and face estimation.

Since marker-based 3D mocap labels are hard to obtain, there are a few annotation methods [13,30,33,37,38,40] for high-precision labeling for both monocular indoor and outdoor scenes. FBA [40] emphasizes the severe failure cases of existing body recovery methods on consumer video data due to unusual camera viewpoints and aggressive truncations. They annotate pseudo 2D body keypoints and SMPL annotations via HMR [21] on 13k frames across four action recognition datasets. Multi-shot-AVA [38] also argues that data from edited media, like movies with rich appearances, interactions between humans, and various temporal contexts, is valuable. They apply the proposed multi-shot optimization on AVA [14] to get pseudo 3D ground truth. Interestingly, a body recovery benchmark [34] finds that simply using the 2D COCO dataset with pseudo-3D labels can surprisingly achieve a better performance and generalization ability. To complement these prior datasets and focus on expressive body recovery, we construct a new benchmark with high-quality 2D and 3D whole-body annotations.

3. Method

3.1. Motivation

A one-stage framework is vital to simplify the cumbersome processes without hand-craft and complex integration designs. However, translating from multi-stage methods directly to a one-stage method is nontrivial. We take the present state-of-the-art method Hand4Whole [29] as an example to perform some preliminary studies on bringing the gap between the multi-stage method and one-stage approach. On the one hand, we replace its separate backbones with a shared backbone for all human components. On the other hand, we explore different crop-and-resize image resolutions for the hands and face, as they usually have small

Type	Dataset	#Frames	Scenes	Multi Person	In-the-wild	Upper Body	Video	Annotation Type	Annotation Source
Rendered	AGORA [35]	17K	Daily	Y	N	N	N	SMPL-X	[35]
Marker/Sensor-based MoCap	Human3.6M [17]	3.6M	Daily	N	N	N	Y	SMPL-X	[30]
	3DPW [45]	> 51K	Daily	Y	Y	N	Y	SMPL-X	[30]
Marker-less Multi-view MoCap	MPI-INF-3DHP [28]	> 1.3M	Daily	N	Y	N	Y	SMPL-X	[30]
	EHF [37]	0.1K	Daily	N	N	N	N	SMPL-X	[37]
	ZJU-MoCap [39]	≥ 237K	Daily	N	N	N	Y	SMPL-X	[1]
Pseudo-3D Labels	PennAction [53]	77K	Fitness	N	Y	N	Y	SMPL	[52]
	MSCOCO [27]	200K	Daily	Y	Y	N	N	SMPL-X	[30]
	COCO-Wholebody [18]	200K	Daily	Y	Y	N	N	2D KPT	[18]
	MPII [3]	25K	Daily	Y	Y	N	N	SMPL-X	[30]
	MTP [32]	3.8K	Daily	N	Y	N	N	SMPL-X	[32]
	FBA [40]	13K	Vlog&Cook&Daily	Y	Y	N	Y	SMPL	[40]
	Multi-shot-AVA [38]	350K	Movie	Y	Y	N	Y	SMPL	[38]
	***UBody* (Ours)**	>1051K	Real-life Scenes	Y	Y	Y	Y	SMPL-X&2D KPT	Ours

Table 1. Comparison of related datasets. *UBody* is a large-scale upper-body dataset with high-precision whole-body annotations.

Method	**AGORA-val**			**EHF**		
	Hand	**Face**	**All**	**Hand**	**Face**	**All**
Ori.	<u>73.3</u>	<u>81.4</u>	<u>183.8</u>	<u>42.7</u>	25.7	77.5
Ori.+1/4 Hand	75.7	**80.8**	**183.0**	50.9	24.8	78.5
Ori.+1/4 Face	**73.2**	81.8	184.0	**41.3**	<u>24.2</u>	**77.0**
Share Backbone	81.1	91.0	202.3	55.5	33.5	84.7
Share+1/4 Hand	77.4	86.1	188.6	57.0	25.8	84.8
Share+1/4 Face	79.5	85.0	196.7	57.8	**24.4**	82.5

Table 2. A preliminary study on the effect of different component scales and share backbone for all components' feature extraction.

image resolutions.

Table 2 shows that, when we transition from the original setup (*Ori.*) to a shared backbone (*Share Backbone*), all recovery errors are severely deteriorated on two datasets. Specifically, MPVPE increases from 183.8mm to 202.3mm (a **10.1**% drop) on AGORA [35], and from 77.5mm to 84.7mm (a **9.3**% drop) on EHF for all components (*All*). These results indicate that extracting the multi-component whole-body features with a shared backbone is difficult. Notably, the hand estimation performance deteriorates by **30.0**% on EHF. Based on the results of different resolutions, we summarize some interesting observations as follows: (i) Overall, changing the resolution of the hand results in a larger performance drop than the face on EHF; (ii) When not sharing a backbone, the results are generally worse with smaller input resolutions of the hands and face.

3.2. Building Component Aware Transformer

As an attempt to break the above status quo, we propose a one-stage framework with a vision transformer encoder and decoder for expressive full-body mesh recovery, named *OSX*. It is simple in design and effective in full-body mesh prediction, as we will demonstrate later. We hope it can serve as a baseline for future one-stage methods. Given a human image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, our component-aware Transformer (*CAT*) estimates the corresponding body, hand, and face parameters $\hat{\mathcal{P}} = \{\hat{\mathbf{P}}_{body}, \hat{\mathbf{P}}_{lhand}, \hat{\mathbf{P}}_{rhand}, \hat{\mathbf{P}}_{face}\}$ and then feed them into a SMPL-X layer [37] to obtain the final 3D whole-body human mesh. Specifically, $\hat{\mathbf{P}}_{body}$ contains 3D body joint rotation $\theta_{body} \in \mathbb{R}^{22\times 3}$, body shape $\beta \in \mathbb{R}^{10}$, and 3D global translation $t \in \mathbb{R}^3$. For $\hat{\mathbf{P}}_{lhand}$ and $\hat{\mathbf{P}}_{rhand}$, they have 3D left and right hand joint rotation $\theta_{lhand} \in \mathbb{R}^{15\times 3}$ and $\theta_{rhand} \in \mathbb{R}^{15\times 3}$, respectively. $\hat{\mathbf{P}}_{face}$ consists of 3D jaw rotation $\theta_{face} \in \mathbb{R}^3$ and facial expression $\phi \in \mathbb{R}^{10}$. Our training target is to minimize the distance between the recovered parameters $\hat{\mathcal{P}}$ and the ground-truth parameters $\mathcal{P}$. As shown in Figure 3, the proposed *CAT* consists of a component-aware encoder to capture the global correlation and extract high-quality multi-scale feature, and a component-aware decoder to strengthen the hand and face regression via an up-sampling strategy to obtain higher-resolution feature maps.

3.3. Body Regression via Global Encoder

In the component-aware encoder, the human image $\mathbf{I}$ is split into fixed-size image patches $\mathbf{P} \in \mathbb{R}^{\frac{HW}{M^2}\times(M^2\times 3)}$, where M is the patch size. The patches $\mathbf{P}$ are then linearly projected by a convolution layer and added with position embeddings $\mathbf{P_e} \in \mathbb{R}^{\frac{HW}{M^2}\times C}$ to obtain a sequence of feature tokens $\mathbf{T_f} \in \mathbb{R}^{\frac{HW}{M^2}\times C}$. To explicitly leverage the body prior and learn the body information in the encoder, we concatenate the feature token $\mathbf{T_f}$ with the body tokens $\mathbf{T_b} \in \mathbb{R}^{B\times C}$, which are learnable parameters. The concatenated tokens are then fed into a standard Transformer encoder with multiple Transformer blocks [11]. Each block consists of a multi-head self-attention, a feed-forward network (FFN), and two layer normalization. After the global feature fusion, the body tokens and image feature tokens are updated into $\mathbf{T_b}' \in \mathbb{R}^{B\times C}$ and $\mathbf{T_f}' \in \mathbb{R}^{\frac{HW}{M^2}\times C}$. Finally, we use several fully connected layers to regress the

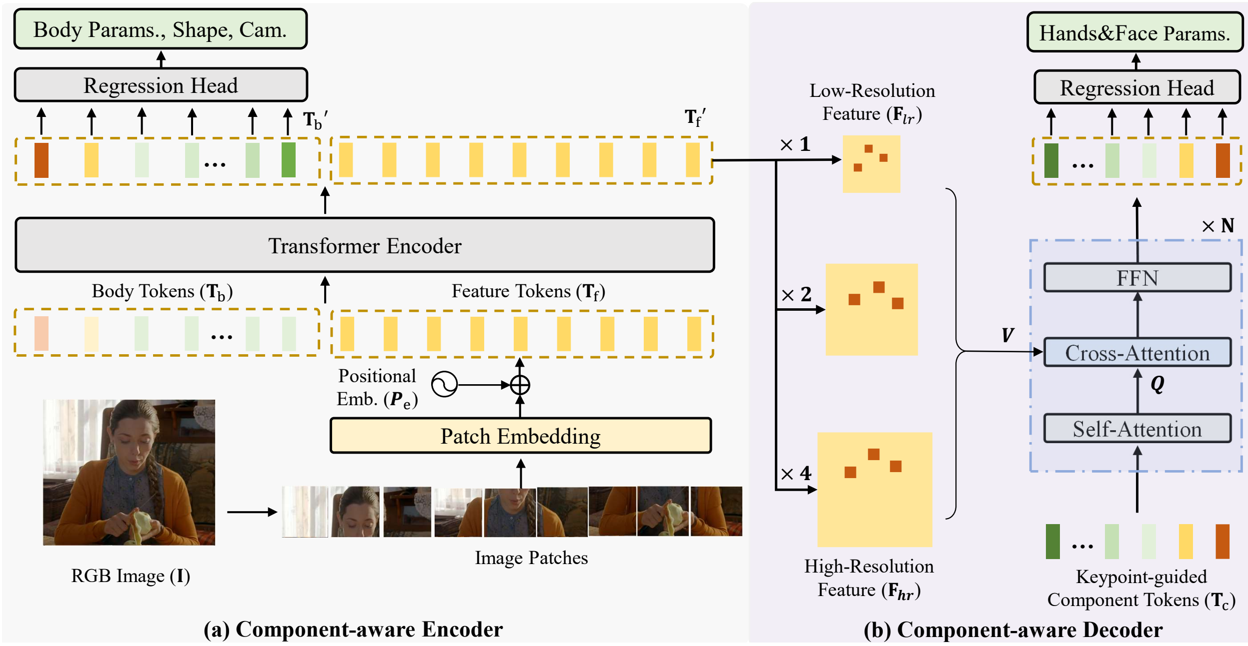


Figure 3. The overview of the proposed one-stage framework (*OSX*) with component-aware transformer. It includes (a) a component-aware Transformer encoder and (b) a component-aware Transformer decoder.

body parameters $\hat{\mathbf{P}}_{body} = \{\theta_{body}, \beta, t\}$ based on $\mathbf{T_b}'$.

3.4. High-Resolution Decoder for Hand and Face

Up-sampling for multi-scale high-resolution features. Since the hands and face in a human image are usually small, previous methods upsample the human image and crop out the hands and face to obtain higher-resolution images. However, this image-level upsampling-crop scheme requires additional backbones to extract the hand and face features separately. To solve this problem, we propose a differentiable feature-level upsampling-crop strategy to enhance the hands and face regression process as inspired by the recent ViTDet [26]. Specifically, we reshape the feature tokens $\mathbf{T_f}'$ into a feature map and upsample it into multiple higher-resolution features $\mathbf{T}_{hr}$ via deconvolution layers. Then, since decoding the hand and face component information from the full feature map inevitably leads to redundant computation and makes the computation process inefficient, we perform differentiable RoIAlign [15] on the feature maps and crop out multi-scale hand feature maps $\mathbf{T}_{hand}$ and face feature maps $\mathbf{T}_{face}$, according to the predicted hand and face bounding boxes, which are regressed from $\mathbf{T_f}'$ using FFNs. The up-sampling and decoding processes for hand and face components are the same, and we illustrate the case of hand parameter regression in detail in Figure 3(b). The cropped multi-scale hand features can be represented as $\mathbf{T}_{hand} = \{\mathbf{F}_{lr}, ..., \mathbf{F}_{hr}\}$. The low-resolution feature $\mathbf{F}_{lr} \in \mathbb{R}^{\frac{H'}{M} \times \frac{W'}{M} \times C}$ is cropped from the original low-resolution feature map, where H' and W' are the height and width of hand image patches. $\mathbf{F}_{hr}$ is the highest-resolution feature. The cropped multi-scale features then serve as memory tokens $\mathbf{V}$ for the keypoint-guided component-aware decoder. To relieve the computational pressure, we reduce the token dimension from C to C' in the component-aware decoder, where $C' = C/2$.

Keypoint-guided deformable attention decoder. To improve the precision of hand and face parameter regression, we leverage 2D keypoint positions as prior knowledge to obtain better component tokens $\mathbf{T_c}$ than random initialization. We simply use the feature map $\mathbf{F}_{lr}$ to regress each 2D keypoint to trade off accuracy and efficiency and regard it as a reference keypoint. The input $\mathbf{T_c} \in \mathbb{R}^{K \times C'}$ of the decoder, which we call the keypoint-guided component tokens, is obtained by summing up reference keypoint feature, pose positional embedding, and learnable embeddings. We then pass the keypoint-guided component token through N deformable attention blocks as inspired by deformable DETR [55]. To relieve the issue of looking over all possible spatial locations, these blocks learn a small set of sampling points (*e.g.*, four here) around the reference keypoint and further enlarge the feature spatial resolution while maintaining computational efficiency compared to vanilla DETR [7]. Each block is composed of a multi-head self-attention layer, a multi-scale deformable cross-attention layer, and FFNs. In the deformable cross-attention layer, keypoint queries $\mathbf{Q}$ extract features from the elements of multi-scale features $\mathbf{V}$ around the position of keypoints p_q:

$$\mathrm{CA}(\mathbf{Q}, \mathbf{V}, p_q) = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{lqk} W \mathbf{V}_l(\phi_l(p_q) + \Delta p_{lqk}), \quad (1)$$

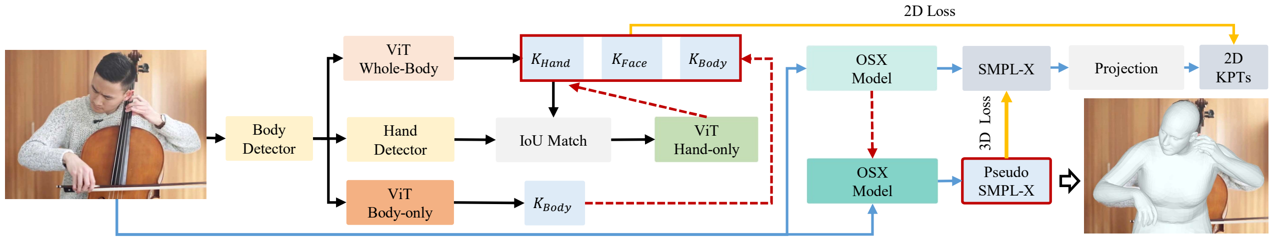


Figure 4. Illustration of the annotation pipeline of *UBody*. Black lines show the annotation process of 2D whole-body keypoints, and blue lines are the 3D SMPL-X annotation procedure. Red dotted lines mean to update the information.

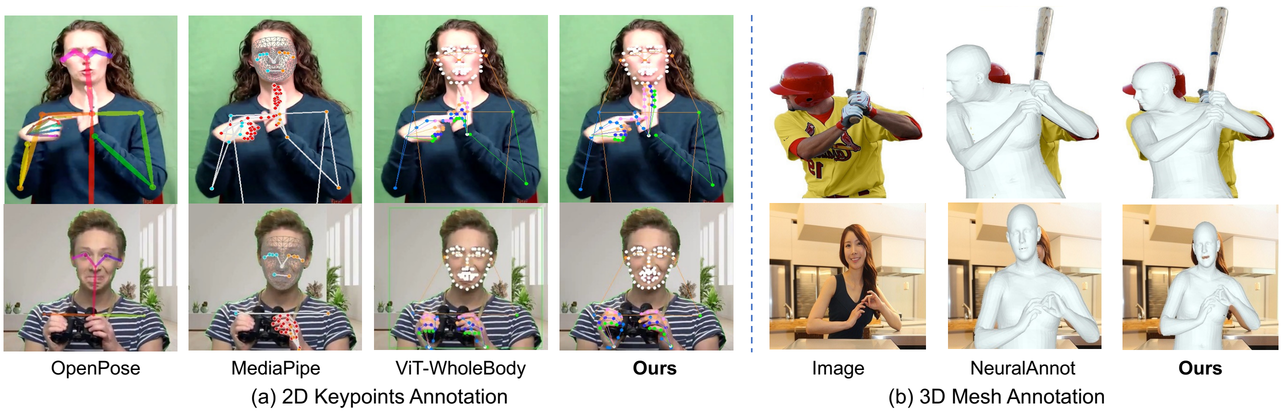


Figure 5. Comparisons of (a) the 2D keypoints annotation quality of wildly used methods [6, 50] and recent SOTA [47] on *UBody* (the left part), and (b) the 3D mesh annotation quality of previous SOTA [30] with ours on COCO (the right part).

where l and k index the feature level and keys, A and W are attention weight and learnable parameter. $\phi(\cdot)$ and Δp are position rescaling and offset. After that, the updated component tokens $\mathbf{T_c}' \in \mathbb{R}^{K \times C'}$ will be fed into hand or face regression head to output the final hand or face parameters $(\hat{\mathbf{P}}_{lhand}, \hat{\mathbf{P}}_{rhand}, \hat{\mathbf{P}}_{face})$, respectively.

Loss Function. *OSX* is trained in an end-to-end manner by minimizing the following loss function:

$$L = L_{smplx} + L_{kpt3D} + L_{kpt2D} + L_{bbox2D}. \quad (2)$$

The four items are calculated as the L1 distance between the ground truth values and the predicted ones. Specifically, L_{smplx} provides the explicit supervision of the SMPL-X parameters. L_{kpt3D}, L_{kpt2D}, and L_{bbox2D} are regression losses for 3D whole-body keypoints, projected 2D whole-body keypoints, and left/right hands and face 2D bounding boxes. More details are provided in the Appendix.

4. UBody–An Upper Body Dataset

3D whole-body mesh recovery from videos is a basic computer vision task, where it can provide comprehensive motion, gesture, and expression information to understand how humans perceive and act. However, existing datasets lack scenes of downstream tasks, such as sign language recognition, gesture generation, emotion recognition, and real-life scenarios recorded as VLOGs, making recent state-of-the-art methods hard to generalize well on these scenes. Interestingly, these scenarios are more concerned with the representations of *upper bodies*. We take this insight and present a novel large-scale benchmark for the expressive *upper body* mesh recovery as shown in Figure 2(f) to (t), named *UBody*. Our annotation pipeline is in Figure 4. *Due to the page limit, we put the data collection, data annotation processes, and annotation visualization in Appendix.*

4.1. Quality Analysis

Our annotation pipeline produces far better 3D pseudo-GT fits with a shorter running time than the previous optimization-based and learning-based methods [13, 20, 30, 37, 38]. Figure 5(a) compares our 2D annotation results with the two wildly used annotation methods (OpenPose [6] and MediaPipe [50]). The quality of our 2D annotations is much more accurate, especially in terms of hand details and the robustness of occlusion and blur. Figure 5(b) compares the 3D annotation of ours with the SOTA NeuralAnnot method [30] on COCO. The quality of our approach is also better for the naked eye in terms of the fit of the body shape and the whole-body poses.

4.2. Data Characteristics

Compared to the popular datasets illustrated in Figure 2 (a) to (e) and the related human-centric datasets listed in Table 1, *UBody* possesses unique features that present new

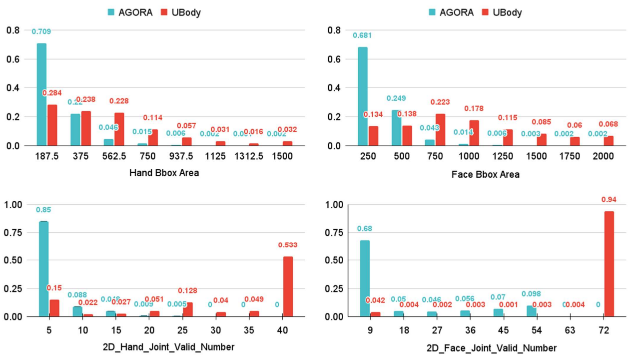


Figure 6. The statistical comparisons of the areas of the hand and face bounding boxes (upper row) and the number of 2D visible hand and face keypoints (lower row) with the logarithmic scale of the Y-axis. *UBody* focuses on upper bodies exhibiting expressive gestures and facial expressions.

challenges for future research. Many videos are from edited media with highly diverse scenes and rich human actions and gestures. They have abrupt shot changes and dynamic camera viewpoints, leading to discontinuities between the frames. Close-up shots of humans cause severe truncation, making existing methods tend to fail. Meanwhile, they have varying degrees of interaction with objects and body components, subtitles, and special effects as occluded scenes. Also, there are high variations in background and light. Those conditions have not appeared in previous datasets. All scenes in *UBody* have rich hand gestures and facial expressions, making the recognition models pay more attention to these important body components. Lastly, all of these real-life videos provide audio as additional information to serve future multi-modality methods. We also provide statistical comparisons between the key features of *UBody* and the wildly used dataset AGORA [35] in Figure 6. AGORA's hand/face bounding box area is generally small, while *UBody* pays more attention to diverse hand and face scales as evidenced by its more dispersed area distribution. Meanwhile, *UBody* has more visible face/hand keypoints, underscoring the importance of recognizing hand gestures and facial expressions. Lastly, *UBody*'s inclusion of real-life videos provides new possibilities for subsequent spatio-temporal modeling that are not available in AGORA, which is an image-based dataset.

5. Experiment

5.1. Experimental Setup

Due to the page limit, we leave the detailed experiment setup, implementation, annotation visualization, qualitative comparison with SOTA methods, and more benchmark results and analyses in the appendix.

Datasets. We use COCO-Wholebody [18], MPII [3], and Human3.6M [17] as the training set. Unlike previous multi-stage methods [29, 36], we do not use additional hand-only and face-only datasets for training as a simple baseline for a one-stage method. The SMPL/SMPL-X pseudo-GTs are obtained from EFT [19] and NeuralAnnot [30].

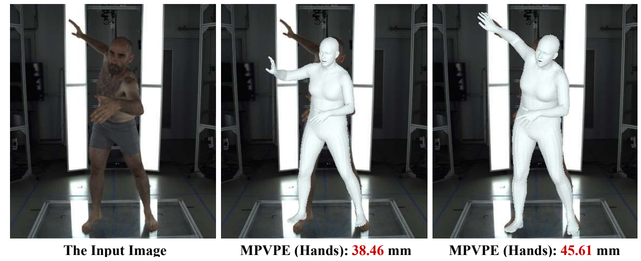


Figure 7. Illustration of the inconsistency between quantitative and qualitative results compared Hand4Whole [29] (the middle one) with *OSX* (the right figure) on EHF.

Evaluation metrics. For 3D whole-body mesh recovery, we utilize the mean per-vertex position error (MPVPE) as our primary metric. In addition, we apply *Procrustes Analysis* (PA) to the recovered mesh, and report the PA-MPVPE after rigid alignment. For AGORA, we also report normalized mean vertex error (N-PMVPE) to compensate for missing detection. Hand error is calculated as the mean of the left and right hands. For 3D body-only recovery on 3DPW, we follow previous works [25, 48] to report the mean per joint position error (MPJPE) and PA-MPJPE. All reported errors are in units of millimeters.

Implementation details. *OSX* is implemented in Pytorch and trained using the Adam optimizer with an initial learning rate of 1×10^{-4} for 14 epochs. Scaling, rotation, random horizontal flip, and color jittering are used as data augmentations during training. We set the number of body tokens $\mathbf{T}_b$ and component tokens $\mathbf{T}_c$ to 27 and 92, respectively.

5.2. Comparisons with Existing Methods

Table 3 provides a comprehensive comparison of *OSX* and existing whole-body mesh recovery methods. As the first one-stage method, *OSX* surpasses existing multi-stage models with complex designs in most cases. Notably, *OSX* has not been trained on hand-only and face-only datasets [22, 31, 56]. Our *All MPVPEs* show a 9.5% improvement on AGORA test set and 7.8% improvement on EHF than SOTA [29]. Since AGORA is a more complex and natural dataset than EHF, previous works [29, 35] claim it is more convincing and representative of real-world scenarios. We also visualize the misleading *high-error* cases on EHF in Figure 7. Besides, we obtain a SOTA performance on the body-only dataset, 3DPW, with a 13.4% error reduction compared to these whole-body methods. More qualitative results are available in the appendix.

5.3. Ablation Study

Impact of the component-aware decoder. Unlike body-only pose estimation, whole-body mesh recovery requires attention to both the body's posture, which is on a larger

Method	AGORA-test					EHF						3DPW	
	MPVPE ↓			N-MPVPE ↓		MPVPE ↓			PA-MPVPE ↓			MPJPE ↓	PA-MPJPE ↓
	All	Hands	Face	All	Body	All	Hands	Face	All	Hands	Face	Body	Body
ExPose [36]	217.3	73.1	51.1	265.0	184.8	77.1	51.6	35.0	54.5	12.8	5.8	93.4	60.7
FrankMocap [41]	-	55.2	-	-	207.8	107.6	42.8	-	57.5	12.6	-	96.7	61.9
PIXIE [13]	191.8	49.3	50.2	233.9	173.4	89.2	42.8	32.7	55.0	11.1	**4.6**	91.0	61.3
Hand4Whole [29]	-	-	-	-	-	79.2	43.2	**25.0**	53.1	12.1	5.8	-	-
Hand4Whole [29]×	135.5	47.2	41.6	144.1	96.0	76.8	**39.8**	26.1	**50.3**	**10.8**	5.8	86.6	54.4
OSX (Ours)	**122.8**↓9.5%	**45.7**	**36.2**	**130.6**	**85.3**	**70.8**↓7.8%	53.7	26.4	48.7	15.9	6.0	**74.7**↓13.4%	**45.1**

Table 3. 3D body reconstruction error comparisons on three existing datasets. × uses additional hand-only and face-only training datasets.

Hand	Ours	w/o *H.D.*	w/o *K.G*	w/o both
MPVPE	**53.7**	55.3	55.1	56.4
PA-MPVPE	**15.9**	17.7	17.6	18.1

Face	Ours	w/o *F.D.*	w/o *K.G*	w/o both
MPVPE	**26.4**	27.2	26.4	26.8
PA-MPVPE	6.0	5.9	**5.8**	6.0

Upsampling	× 1	× 2	× 4	× 8
MPVPE	54.9	54.3	**53.7**	54.1

Table 4. Ablation study of component-aware decoder on EHF with *H.D.*, *F.D.*, *K.G*, and upsampling strategies. *H.D.*, *F.D.*, and *K.G* are abbreviations for Hand Decoder, Face Decoder and Keypoint-Guided scheme.

Method	MPVPE ↓			PA-MPVPE ↓			PA-MPJPE ↓	
	All	Hand	Face	All	Hand	Face	Body	Hand
ExPose [36]	171.5	83.7	45.1	66.9	12.0	3.9	70.7	12.3
PIXIE [13]	168.4	55.6	45.2	61.7	12.2	4.2	66.8	12.3
Hand4Whole [29]	104.1	45.7	27.0	44.8	8.9	2.8	45.5	9.0
Hand4Whole [29]×	157.4	62.2	49.8	82.2	9.8	3.9	92.8	10.0
OSX (Ours)	92.4	47.7	24.9	42.4	10.8	2.4	42.9	11.0
OSX (Ours)†	**81.9**	**41.5**	**21.2**	**42.2**	**8.6**	**2.0**	**48.4**	**8.8**

Table 5. Reconstruction errors on *UBody* test set on the *intra-scene* protocol. All models are pretrained on previous datasets, except for the results labeled by (i) †: finetuned on the *UBody* training data; (ii) ×: finetuned on the AGORA training data. The result of the *inter-scene* setting is in the appendix.

spatial scale, and the gesture and expression of the hands and face, which are on a finer scale. To handle the resolution issue in a one-stage pipeline, we propose the component-aware decoder attached to the component-aware encoder. First, in the upper Table 4, we verify the effectiveness of the proposed decoder for both hand and face regression. We can observe a significant drop without the decoder (*e.g.*, *w/o H.D.* and *w/o F.D.*, indicating that simply regressing the low-resolution hand and face directly from the encoder is inferior. Moreover, the errors will also increase without the proposed keypoint-guided deformable attention scheme, as shown in the medium Table 4. In particular, the performance of the hand estimation is highly influenced, showing that hand pose estimation attends more to the sparsely deformable spatial information to obtain better queries.

Impact of the up-sampling strategy. To relieve the low-resolution problem of hand and facial features, we design the feature up-sampling strategy in the decoder to obtain multi-scale higher-resolution features. The lower Table 4 presents the impact of different up-sampling scale. As the up-sampling scale increases, the MPVPE decreases and then reaches a saturation point. Therefore, we use three scales (i.e., $[\times 1, \times 2, \times 4]$) by default in our experiments.

5.4. Benchmark on UBody

As a new dataset, we provide both quantitative and qualitative results on *UBody*. Table 5 presents the performance comparisons of existing 3D whole-body methods. The general result ranking is similar to AGORA. Since the upper body is closer to the camera, their errors will be smaller than AGORA. However, the hand and face will play a more important role than previous data. Besides, we finetune Hand4Whole on AGORA and test again, and we find all errors are significantly enlarged. This observation can be attributed to the data distribution gap between AGORA and *UBody*, as shown in Figure 6. Moreover, we train *OSX* on our train set and find a 16.1% improvement compared to the original pretrained model, indicating that *UBody* can serve to improve the performance on downstream real-life scenes.

6. Conclusion

In this work, we propose the first one-stage pipeline for 3D whole-body mesh recovery that achieves SOTA performance on three benchmarks in a simple yet effective manner. Moreover, to bridge the gap between the basic task of full-body pose and shape estimation and their downstream tasks, we develop a large-scale dataset with comprehensive scenes covering our daily life. With our proposed annotation method, we show that training on *UBody* can effectively improve the performance of mesh recovery in upper-body scenes. We hope this work can contribute new insights to this area, both in terms of methodology and dataset.

Limitation and future work. Currently, our training does not use additional hand and face-specific datasets. It is worth studying how to make the best use of them in our pipeline to further improve performance. Also, we can validate the effectiveness of *UBody* on some downstream applications, *e.g.*, gesture recognition, driving avatar.

Acknowledgements: This work was partially funded through the National Key Research and Development Program of China (Project No.2022YFB36066), in part by the Shenzhen Science and Technology Project under Grant (CJGJZD20200617102601004, JCYJ20220818101001004).

References

[1] Easymocap - make human motion capture easier. Github, 2021. 4

[2] Oswald Aldrian and William A. P. Smith. Inverse rendering of faces with a 3d morphable model. *TPAMI*, 2013. 1, 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 4, 7

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[5] 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 *ECCV*, 2022. 3

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2021. 6

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5

[8] Theocharis Chatzis, Andreas Stergioulas, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 2020. 2

[9] 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 *ECCV*. 2020. 1

[10] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4

[12] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models—past, present, and future. *TOG*, 2020. 1, 2

[13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 2, 3, 6, 8

[14] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 3

[15] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5

[16] Lin Huang, Zhang Boshen, Zhilin Guo, Yang Xiao, Zhiguo Cao, and Junsong Yuan. Survey on depth and rgb image-based 3d hand shape and pose estimation. *Virtual Reality & Intelligent Hardware*, 2021. 2

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 3, 4, 7

[18] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 4, 7

[19] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. 2020. 7

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *3DV*, 2022. 3, 6

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *TPAMI*, 2021. 7

[23] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 2

[24] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *CVPR*, 2021. 2

[25] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 2, 7

[26] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 5

[27] 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 *ECCV*, 2014. 4

[28] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV, year=2017*. 4

[29] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, 2022. 2, 3, 7, 8

[30] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3d human mesh training sets. In *CVPRW*, 2022. 3, 4, 6, 7

[31] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *ECCV*, 2020. 7

[32] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 4

[33] Lea Muller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *CVPR*, 2021. 3

[34] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Neural Information Processing Systems Datasets and Benchmarks Track*. 3

[35] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 2, 3, 4, 7

[36] Georgios Pavlakos, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, Michael J. Black, Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. *ECCV*, 2020. 1, 2, 7, 8

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 3, 4, 6

[38] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 3, 4, 6

[39] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 4

[40] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 3, 4

[41] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 2, 8

[42] Yu Sun, Tianyu Huang, Qian Bao, Wu Liu, Wenpeng Gao, and Yili Fu. Learning monocular mesh recovery of multiple body parts via synthesis. In *ICASSP*. IEEE, 2022. 2

[43] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCVW*, 2017. 1, 2

[44] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv:2203.01923*, 2022. 1, 2

[45] 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 *ECCV*. 2018. 3, 4

[46] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 2

[47] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv:2204.12484*, 2022. 2, 6

[48] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *ECCV*, 2022. 1, 2, 7

[49] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *ECCV*, 2022. 1

[50] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv:2006.10214*, 2020. 6

[51] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *arXiv:2207.06400*, 2022. 2

[52] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3d human dynamics from video. In *ICCV*, 2019. 4

[53] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 4

[54] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular realtime full body capture with inter-part correlations. In *CVPR*, 2021. 2

[55] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021. 5

[56] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 7