

# HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu<sup>†</sup>, *Member, IEEE*

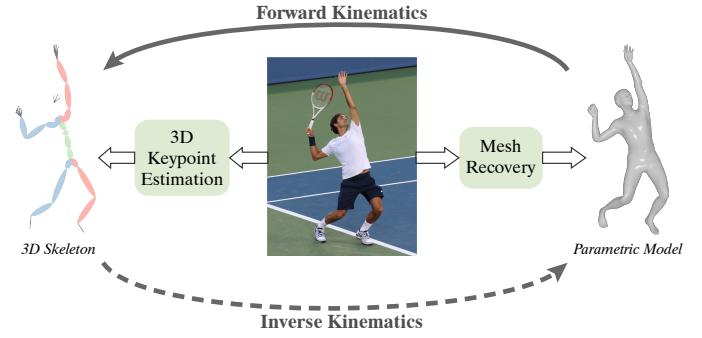
**Abstract**—Recovering whole-body mesh by inferring the abstract pose and shape parameters from visual content can obtain 3D bodies with realistic structures. However, the inferring process is highly non-linear and suffers from image-mesh misalignment, resulting in inaccurate reconstruction. In contrast, 3D keypoint estimation methods utilize the volumetric representation to achieve pixel-level accuracy but may predict unrealistic body structures. To address these issues, this paper presents a novel hybrid inverse kinematics solution, HybrIK, that integrates the merits of 3D keypoint estimation and body mesh recovery in a unified framework. HybrIK directly transforms accurate 3D joints to body-part rotations via twist-and-swing decomposition. The swing rotations are analytically solved with 3D joints, while the twist rotations are derived from visual cues through neural networks. To capture comprehensive whole-body details, we further develop a holistic framework, HybrIK-X, which enhances HybrIK with articulated hands and an expressive face. HybrIK-X is fast and accurate by solving the whole-body pose with a one-stage model. Experiments demonstrate that HybrIK and HybrIK-X preserve both the accuracy of 3D joints and the realistic structure of the parametric human model, leading to pixel-aligned whole-body mesh recovery. The proposed method significantly surpasses the state-of-the-art methods on various benchmarks for body-only, hand-only, and whole-body scenarios. Code and results can be found at <https://jeffli.site/HybrIK-X/>.

**Index Terms**—hybrid inverse kinematics solution, whole-body mesh recovery, 3D human from monocular images.

## 1 INTRODUCTION

RECOVERING the whole-body 3D surface from visual content has a broad spectrum of applications. Advancements in parametric statistical human body shape models [1], [2], [3] have enabled the generation of realistic and animatable 3D meshes using only a small set of parameters. Despite the recent progress, recovering the 3D mesh by inferring the abstract pose and shape parameters is highly non-linear and still challenging.

Existing approaches can be divided into two categories: optimization-based and learning-based. Optimization-based approaches [3], [4], [5] estimate the pose and shape of the human body through an iterative fitting process. The parameters of the statistical model are optimized to reduce the error between its 2D projection and 2D observations, e.g., 2D joint positions and silhouettes. However, this optimization problem is non-convex. Its solution can be time-consuming and its results are sensitive to the initialization. These challenges have shifted the research focus towards learning-based approaches. Learning-based approaches leverage parametric body models and employ neural networks to directly regress the pose parameters [6], [7], [8], [9], [10]. Nevertheless, the pose parameters represent relative rotations that lie in the  $\mathcal{SO}(3)$  rotation group, which puts difficulties for neural networks to learn directly from RGB images. Consequently, the learned body mesh suffers from image-mesh misalignment.



**Fig. 1. Closing the loop between the 3D skeleton and the parametric model via HybrIK/HybrIK-X.** A 3D skeleton predicted by the neural network can be transformed into a parametric body mesh by inverse kinematics (IK) without loss of accuracy. The parametric body mesh can generate structural realistic 3D skeleton by forward kinematics (FK).

Such a challenge prompts us to look into the field of 3D keypoint estimation. Previous 3D keypoint estimation approaches [11], [12], [13] adopt volumetric heatmap as the target representation to learn 3D joint positions in the Cartesian coordinate system. The learned 3D joints can accurately align with the 2D RGB image. This inspires us to establish a collaboration between the 3D joints and the body mesh via forward kinematics (FK) and inverse kinematics (IK) (as illustrated in Fig. 1). On the one hand, the accurate 3D joints can improve image-mesh alignment for mesh recovery. Since the body mesh is recovered from 3D joints, the recovered mesh can obtain pixel-aligned accuracy as long as the 3D joints are well-aligned with the image. On the other hand, the shape prior inherent in the parametric body

- Jiefeng Li, Siyuan Bian, Zhicun Chen, Lixin Yang and Cewu Lu are with the Department of Electrical and Computer Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. E-mail: {jlf\_likit, biansiyuan, zhicun\_chen, siriusyang, lucewu}@sjtu.edu.cn.
- Chao Xu is with Xiaoice Company, Shanghai, China. E-mail: {xuchao.19962007}@sjtu.edu.cn.
- Corresponding author: Cewu Lu.

model can be utilized to mitigate the issue of the unrealistic body structure in 3D keypoint estimation approaches. Since existing 3D keypoint estimation approaches lack explicit modeling of body bone length, they may predict unrealistic body structures like left-right asymmetry and abnormal proportions of limbs. If we can leverage the parametric body model, the presented human shape can better conform to the actual human body.

In this work, we present a hybrid analytical-neural inverse kinematics solution (HybRIK) to establish a collaboration between 3D keypoint estimation and whole-body mesh estimation. Inverse kinematics (IK) is used to find the corresponding body-part rotations from 3D body joints. This is an ill-posed problem due to the lack of a unique solution. The core of our approach is an innovative solution for this problem using twist-and-swing decomposition. Specifically, the rotation of a skeleton part is decomposed into *twist* and *swing*, i.e., a longitudinal rotation and an in-plane rotation. The unique solution of the body-part rotations is composed iteratively along the kinematic tree by analytically calculating *swing* rotations from 3D joints and using a neural network to predict *twist* rotations from visual cues.

This IK framework is extended as HybRIK-X for articulated hand and expressive face reconstruction, improving the fine-grained image-mesh alignment with a new backward-updated solution. Unlike previous approaches [10], [14], [15], HybRIK-X gets rid of separated expert models and recovers the whole-body mesh with a one-stage network, resulting in improved efficiency and reduced computational resources. The robustness of 3D keypoints estimation against occlusions and truncations is enhanced by using a new regression approach to infer the truncated body parts. Furthermore, we exploit the body structure from the parametric human body model to alleviate the depth ambiguity in predicting camera parameters and estimate a more stable human motion.

A critical characteristic of our approach is that the estimated mesh is inherently aligned with the 3D skeleton, without the need for additional optimization procedures in the previous approaches [3], [5], [7]. We conduct comprehensive experiments on various benchmarks for body-only, hand-only, and whole-body scenarios, including 3DPW [16], Human3.6M [17], and MPI-INF-3DHP [18], FreiHAND [19], HO3D [20], and AGORA [21]. HybRIK and HybRIK-X show pixel-aligned accuracy and significantly outperform state-of-the-art approaches.

The contributions of our approach can be summarized as follows:

- We propose a novel whole-body mesh recovery framework that uses a hybrid analytical-neural IK algorithm to convert accurate 3D joints to pixel-aligned body meshes.
- Our approach closes the loop between the 3D skeleton and the parametric model. It improves the image-mesh alignment for body mesh recovery and addresses the unrealistic body structure problem of 3D keypoint estimation approaches at the same time.
- Our approach achieves state-of-the-art performance across various body-only, hand-only, and whole-body benchmarks.

A preliminary version of this work was accepted in CVPR 2021 [22]. This paper extends the previous work in the following ways. First, we extend the IK framework to whole-body mesh recovery with a backward-updated IK solution and an efficient one-stage model. Second, we enhance the robustness of our framework to occlusions and truncations by a new regression paradigm. Third, we propose a structure-aware cycle for the mitigation of depth ambiguity, thereby providing a more stable camera parameters estimation. Fourth, additional quantitative comparisons on various benchmarks for body-only, hand-only, and whole-body scenarios demonstrate the effectiveness and generalization of the proposed IK framework. Finally, we conduct further ablation studies to investigate and analyze our framework.

## 2 RELATED WORK

### 2.1 3D Keypoint Estimation

Many studies formulate 3D human pose estimation as the problem of locating the 3D joints of the human body. Previous work can be divided into two categories: single-stage and two-stage approaches. Single-stage approaches [13], [23], [24], [25], [26], [27], [28], [29] directly estimate the 3D joint locations from the input image. Various representations are developed, including 3D heatmap [23], location-map [25], and 2D heatmap +  $z$  regression [26]. Two-stage approaches first estimate 2D pose and then lift them to 3D joint locations by a learned dictionary of 3D skeleton [30], [31], [32], [33], [34], [35] or regression [36], [37], [38], [39], [40], [41]. Two-stage approaches highly rely on accurate 2D pose estimators, which have achieved impressive performance through the combination of a powerful backbone network [42], [43], [44], [45], [46] and the 2D heatmap.

These privileged forms of supervision contribute to the recent performance leaps of 3D keypoint estimation. However, the human structural information is modeled implicitly by the neural network, which can not ensure the output 3D skeletons are realistic. Our approach combines the advantages of both the 3D skeleton and parametric model to predict accurate and realistic human pose and shape.

### 2.2 Model-based 3D Body Pose and Shape Estimation

Prior work on the model-based 3D pose and shape estimation uses parameters of the statistical body model [1], [2], [3] as the output target because they capture the statistics prior of body shape. Compared with the model-free methods [47], [48], [49], the model-based methods directly predict controllable body mesh, which can facilitate many downstream tasks for both computer graphics and computer vision. Bogo et al. [5] propose SMPLify, a fully automatic approach, without manual user intervention [4], [50]. This optimization paradigm was further extended with silhouette cues [51], volumetric grids [47], multiple people [52], and whole-body parametric models [3].

With the advances in deep learning networks, there are increasing studies that focus on learning-based methods [6], [7], [8], [9], [53], [54], [55], [56], [57], [58], using a deep network to estimate the pose and shape parameters. Since the mapping from RGB image to body shape and body-part rotation is hard to learn, many studies use intermediate

representations to alleviate this problem, such as keypoints and silhouettes [53], semantic part segmentation [54], and 2D heatmap input [59]. Kanazawa et al. [6] use an adversarial prior and an iterative error feedback (IEF) loop to reduce the difficulty of regression. Arnab et al. [60] and Kocabas et al. [8] exploit temporal context, while Guler et al. [61] use a part-voting expression and test-time post-processing to improve the regression network. Kolotouros et al. [7] leverage the optimization paradigm to provide extra 3D supervision from unlabeled images. Zhang et al. [62] propose to use saliency maps to infer occluded bodies. PARE [9] uses part-based attention to improve body-part regression. PyMAF [55] uses a mesh-aligned feedback loop to exploit locally aligned features.

In this work, we address this challenging learning problem by a transformation from the pixel-aligned 3D joints to the body-part rotations.

### 2.3 Whole-body Mesh Recovery

Numerous prior studies solve body, face [63], [64], [65], [66], [67], and hand meshes [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78] separately. Recent studies start using whole-body statistical models [3], [79], [80] to jointly recover the 3D surface of the body, face, and hands.

Pavlakos et al. [3] propose to estimate whole-body mesh by automatically fitting the SMPL-X model [3] to the 2D body, face, and hand keypoints estimated by off-the-shelf whole-body keypoint estimators [81], [82]. Xiang et al. [83] propose to learn 2D keypoints and part orientation fields for model fitting. Xu et al. [80] fit GHUM with a reprojection error and a semantic body-part alignment error under anatomical joint angle limits. Similar to body-only pose estimation, optimization-based methods are slow and sensitive to initialization.

Learning-based methods tackle the above limitations by directly regressing the pose and shape parameters from the input image. ExPose [84] uses three expert sub-networks to estimate body, face, and hands parameters separately. The body expert first estimates the body pose and the rough poses for the face and hands. Then the face and hand experts estimate the refined part poses from the cropped local image. Finally, the whole-body mesh is reconstructed by merging the results from three experts. Follow-up studies follow the same paradigm of ExPose [84]. They use separate experts to handle body, face, and hand poses and merge them into a holistic whole-body pose. FrankMocap [14] integrates the results from three experts through an integration network to approximate the optimization to better align the wrist and hand poses. Zhou et al. [85] propose to regress the body and hand poses from the detected keypoints. PIXIE [86] uses an attention-based moderator network to integrate body, face, and hand results from experts. Hand4Whole [87] introduces to learn the wrist rotations from the hand keypoints. PyMAF-X [15] proposes an adaptive integration module to better integrate the elbow and wrist poses.

Although using separate experts can benefit network training by using high-resolution input and multiple data sources, it costs higher computational complexity and lacks holistic information to integrate results from different experts. Unlike previous methods, we propose a one-stage

paradigm that directly estimates the whole-body mesh. Our one-stage method gets rid of the time-consuming separated experts and the entire model is trained in an end-to-end manner.

### 2.4 Body-part Rotation in Pose Estimation

The core of our approach is to calculate the relative rotation of human body parts through a hybrid IK process. There are several studies that estimate the rotations in the 3D pose estimation literature. Zhou et al. [88] use the network to predict the rotation angle of each body joint, followed by an FK layer to generate the 3D joint coordinates. Pavllo et al. [89] switch to quaternions, while Yoshiyasu et al. [90] directly predict the  $3 \times 3$  rotation matrices. Mehta et al. [91] first estimate the 3D joints and then use a fitting procedure to find the rotation Euler angles. Previous approaches are either limited to a hard-to-learn problem or require an additional fitting procedure. Our approach recovers the body-part rotation from 3D joint locations in a direct, accurate and feed-forward manner.

### 2.5 Inverse Kinematics Process

The inverse kinematics (IK) problem has been extensively studied during recent decades. Numerical solutions [92], [93], [94], [95], [96], [97] are simple ways to implement the IK process, but they suffer from time-consuming iterative optimization. Heuristic methods are efficient solutions to the IK problem. For example, CDC [98], FABRIK [99], and IK-FA [100] have a low computational cost for each heuristic iteration. In some special cases, there exist analytical solutions to the IK problem. Tolani et al. [101] propose a reliable algorithm by the combination of analytical and numerical methods. Kallmann et al. [102] solve the IK for arm linkage, i.e., a three-joint system. Recently, researchers have been interested in using neural networks to solve the IK problem in robotic control [103], motion retargeting [104], and hand pose estimation [105], [106].

In this work, we combine the interpretable characteristic of the analytical solution and the flexibility of the neural network, introducing a feed-forward hybrid IK algorithm with twist-and-swing decomposition. Twist-and-swing decomposition is first introduced by Baerlocher et al. [107]. The twist angles are limited based on the particular body joint. In our work, the twist angles are estimated by a neural network, which is more flexible and can be generalized to all body joints. Compared with previous analytical solutions [102] designed for specific joint linkage, our algorithm can be applied to the entire body skeleton in a direct and differentiable manner.

## 3 METHOD

In this section, we present our hybrid analytical-neural inverse kinematics solution for whole-body mesh recovery (Fig. 2). First, in §3.1, we briefly describe the forward kinematics process, the inverse kinematics process, and the SMPL/SMPL-X model. In §3.2, we introduce the proposed inverse kinematics solution, HybrIK, for body-only mesh recovery, and HybrIK-X, for whole-body mesh recovery. Then, in §3.3, we present the overall learning framework to

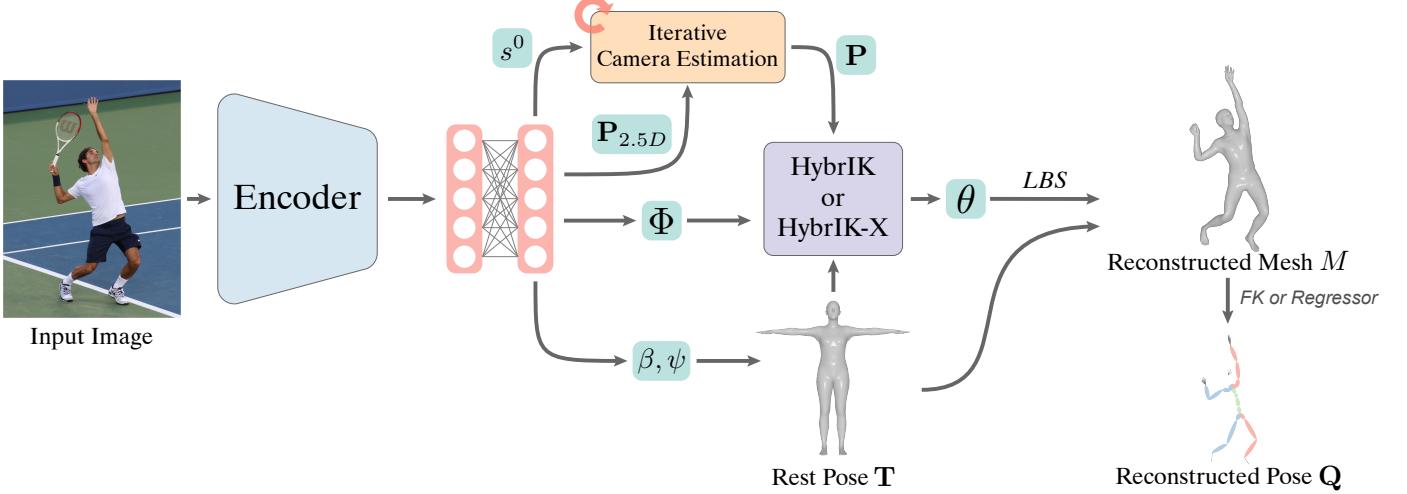


Fig. 2. **Overview of the proposed inverse kinematics framework.** 2.5D joints  $P_{2.5D}$ , shape parameters  $\beta$ , expression parameters  $\psi$ , twist angles  $\Phi$ , and the initial camera parameter  $s^0$  are learned from the visual cues through neural networks. The 2.5D joints  $P_{2.5D}$  and the initial camera parameter  $s^0$  are fed into our iterative camera estimation algorithm to obtain the estimated camera and the back-projected 3D joints  $P$ . These results are then sent to the HybrIK process to solve the body-part rotations, i.e., the pose parameters  $\theta$ . Finally, with the pose and shape parameters, we can obtain the reconstructed body mesh  $M$  via linear blend skinning (LBS), and the reconstructed pose  $Q$  via a further FK process or linear regression.

estimate the pixel-aligned whole-body mesh and realistic 3D skeleton. Finally, we provide the necessary implementation details in §3.4.

### 3.1 Preliminary

**Forward Kinematics.** Forward kinematics (FK) for human pose usually refers to the process of computing the reconstructed pose  $Q = \{q_k\}_{k=1}^K$ , with the rest pose template  $T = \{t_k\}_{k=1}^K$  and the relative rotations  $R = \{R_{pa(k),k}\}_{k=1}^K$  as input:

$$Q = FK(R, T), \quad (1)$$

where  $K$  is the number the body joints,  $q_k \in \mathbb{R}^3$  denotes the reconstructed 3D location of the  $k$ -th joint,  $t_k \in \mathbb{R}^3$  denotes the  $k$ -th joint location of the rest pose template,  $pa(k)$  returns the parent's index of the  $k$ -th joint, and  $R_{pa(k),k}$  is the relative rotation of  $k$ -th joint with respect to its parent joint. FK can be performed by recursively rotating the template body part from the root joint to the leaf joints:

$$q_k = R_k(t_k - t_{pa(k)}) + q_{pa(k)}, \quad (2)$$

where  $R_k \in \mathbb{SO}(3)$  is the global rotation of the  $k$ -th joint with respect to the canonical rest pose space. The global rotation can be calculated recursively:

$$R_k = R_{pa(k)} R_{pa(k),k}. \quad (3)$$

For the root joint that has no parent, we have  $q_0 = t_0$ .

**Inverse Kinematics.** Inverse kinematics (IK) is the reverse process of FK, computing relative rotations  $R$  that can generate the desired locations of input body joints  $P = \{p_k\}_{k=1}^K$ . This process can be formulated as:

$$R = IK(P, T), \quad (4)$$

where  $p_k$  denotes the  $k$ -th joint of the input pose. Ideally, the resulting rotations should satisfy the following condition:

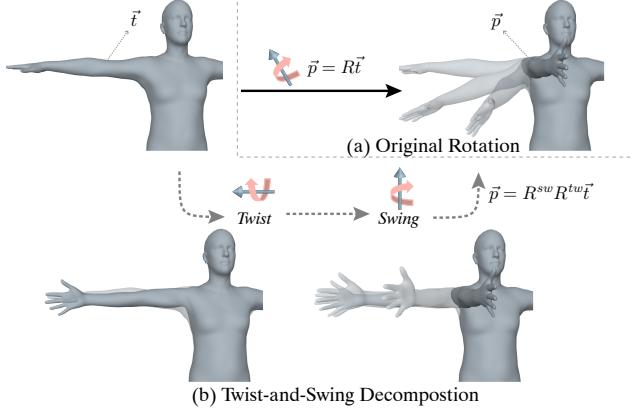
$$p_k - p_{pa(k)} = R_k(t_k - t_{pa(k)}) \quad \forall 1 \leq k \leq K. \quad (5)$$

Similar to the FK process, we have  $p_0 = t_0$  for the root joint that has no parent. While the FK problem is well-posed, the IK problem is ill-posed because there are either no solutions or too many solutions to fulfill the target joint locations.

**SMPL and SMPL-X Models.** In this work, we employ the SMPL [2] parametric model for body-only mesh recovery and the expressive SMPL-X [3] model for whole-body mesh recovery. SMPL-X is the whole-body extension of SMPL. It allows us to use shape parameters, expression parameters and pose parameters to control the whole-body mesh. The shape parameters  $\beta \in \mathbb{R}^{200}$  are parameterized by the first 200 principal components of a linear shape space, learned from registered CAESAR [108] scans. The expression parameters  $\psi \in \mathbb{R}^{50}$  are coefficients of a low-dimensional linear space. The pose parameters  $\theta$  are modelled by relative 3D rotations of  $K = 55$  joints,  $\theta = (\theta_1, \theta_2, \dots, \theta_K)$ , consisting of body, jaw and hand poses. SMPL-X provides a differentiable skinning function  $\mathcal{M}(\theta, \beta, \psi)$  that takes pose parameters  $\theta$ , shape parameters  $\beta$ , and expression parameters  $\psi$  as input and outputs a triangulated mesh  $M \in \mathbb{R}^{N \times 3}$  with  $N = 10475$  vertices. The reconstructed 3D joints  $Q_{smpl-x}$  can be conveniently obtained by the FK process, i.e.,  $Q_{smpl-x} = FK(R, T)$ . Also, other format of joints can be obtained by a linear combination of the mesh vertices through a linear regressor  $W$ , i.e.,  $Q = WM$ .

### 3.2 Hybrid Analytical-Neural Inverse Kinematics

Estimating the human body mesh by directly regressing the body part rotations is too difficult [6], [7], [8]. In this paper, we propose a hybrid analytical-neural inverse kinematics solution that use 3D keypoints to recover 3D body mesh. Since the IK problem is ill-posed, we cannot uniquely determine the relative rotation just by the 3D joints. Here, we first decompose the original rotation into *twist* and *swing*. The 3D joints are utilized to calculate the *swing* rotation



**Fig. 3. Illustration of the twist-and-swing decomposition.** (a) The original rotation moves the right hand to the front and turns the palm to the left in one step. (b) With twist-and-swing decomposition, the rotation can be divided into two steps: First, turn the palm  $90^\circ$  by the *twist* rotation, and then move the entire hand to the front by the *swing* rotation.

analytically, and we exploit the visual cues by a neural network to estimate the 1-DoF *twist* rotation. In our IK algorithm, the body part rotations are solved recursively along the kinematic tree.

### 3.2.1 Twist-and-Swing Decomposition.

In the conventional analytical IK formulation, some body joints are usually assigned lower degree-of-freedom (DoFs) to simplify the problem, e.g., 1 or 2 DoFs [101], [102], [109]. In this work, we consider a general case where each body joint is assumed to have full 3 DoFs. As illustrated in Fig. 3, a rotation  $R \in \mathbb{SO}(3)$  can be decomposed into a *twist* rotation  $R^{tw}$  and a *swing* rotation  $R^{sw}$ . Given the start template body-part vector  $\vec{t}$  and the target vector  $\vec{p}$ , the solution process of  $R$  can be formulated as:

$$R = \mathcal{D}(\vec{p}, \vec{t}, \phi) = \mathcal{D}^{sw}(\vec{p}, \vec{t}) \mathcal{D}^{tw}(\vec{t}, \phi) = R^{sw} R^{tw}, \quad (6)$$

where  $\phi$  is the *twist* angle that estimated by a neural network,  $\mathcal{D}^{sw}(\cdot)$  is a closed-form solution of the *swing* rotation, and  $\mathcal{D}^{tw}(\cdot)$  transforms  $\phi$  to the *twist* rotation. Here,  $R$  should satisfy the condition in Eq. 5, i.e.,  $\vec{p} = R\vec{t}$ .

**Swing Rotation.** The *swing* rotation has the axis  $\vec{n}$  that is perpendicular to  $\vec{t}$  and  $\vec{p}$ . Therefore, it can be formulated as:

$$\vec{n} = \frac{\vec{t} \times \vec{p}}{\|\vec{t}\| \|\vec{p}\|}, \quad (7)$$

and the *swing* angle  $\alpha$  satisfies:

$$\cos \alpha = \frac{\vec{t} \cdot \vec{p}}{\|\vec{t}\| \|\vec{p}\|}, \quad \sin \alpha = \frac{\|\vec{t} \times \vec{p}\|}{\|\vec{t}\| \|\vec{p}\|}. \quad (8)$$

Hence, the closed-form solution of the *swing* rotation  $R^{sw}$  can be derived by the *Rodrigues formula*:

$$R^{sw} = \mathcal{D}^{sw}(\vec{p}, \vec{t}) = \mathcal{I} + \sin \alpha [\vec{n}]_{\times} + (1 - \cos \alpha) [\vec{n}]_{\times}^2, \quad (9)$$

where  $[\vec{n}]_{\times}$  is the skew symmetric matrix of  $\vec{n}$  and  $\mathcal{I}$  is the  $3 \times 3$  identity matrix.

---

### Algorithm 1: Naive HybrIK

---

**Input:**  $P, T, \Phi$

**Output:**  $R$

- 1 Determine  $R_0$ ;
  - 2 **for**  $k$  along the kinematic tree **do**
  - 3      $\vec{p}_k \leftarrow R_{pa(k)}^{-1}(p_k - p_{pa(k)})$ ;
  - 4      $\vec{t}_k \leftarrow (t_k - t_{pa(k)})$ ;
  - 5      $R_{pa(k),k}^{sw} \leftarrow \mathcal{D}^{sw}(\vec{p}_k, \vec{t}_k)$ ;
  - 6      $R_{pa(k),k}^{tw} \leftarrow \mathcal{D}^{tw}(\vec{t}_k, \phi_k)$ ;
  - 7      $R_{pa(k),k} \leftarrow R_{pa(k),k}^{sw} R_{pa(k),k}^{tw}$ ;
- 

**Twist Rotation.** The *twist* rotation is rotating around  $\vec{t}$  itself. Thus, with  $\vec{t}$  itself the axis and  $\phi$  the angle, we can determine *twist* rotation  $R^{tw}$ :

$$R^{tw} = \mathcal{D}^{tw}(\vec{t}, \phi) = \mathcal{I} + \frac{\sin \phi}{\|\vec{t}\|} [\vec{t}]_{\times} + \frac{(1 - \cos \phi)}{\|\vec{t}\|^2} [\vec{t}]_{\times}^2, \quad (10)$$

where  $[\vec{t}]_{\times}$  is the skew symmetric matrix of  $\vec{t}$ .

Note that function  $\mathcal{D}^{sw}$  and  $\mathcal{D}^{tw}$  are fully differentiable, which allows us to integrate the twist-and-swing decomposition into the training process. Although we need a neural network to learn the *twist* angle, the difficulty of learning is significantly reduced. Compared with previous work [6], [7], [8] that directly regresses the 3-DoF rotation, we regress the *twist* angle that is only a 1-DoF variable. Moreover, due to the physical limitation of the human body, the *twist* angle has a small range of variation. Therefore, it is much easier for the networks to learn the mapping function. We further analyze its variation in §4.5.

### 3.2.2 Body-only Inverse Kinematics

**Naive HybrIK.** Using the twist-and-swing decomposition, the IK process can be performed recursively along the kinematic tree like the FK process. First of all, we need to determine the global root rotation  $R_0$ , which has a closed-form solution using the locations of spine, left hip, right hip and Singular Value Decomposition (SVD). Detailed mathematical proof is provided in appendix §A. Then, in each step, e.g., the  $k$ -th step, we assume the rotation of the parent joint  $R_{pa(k)}$  is known. Hence, we can reformulate Eq. 5 with Eq. 3 as:

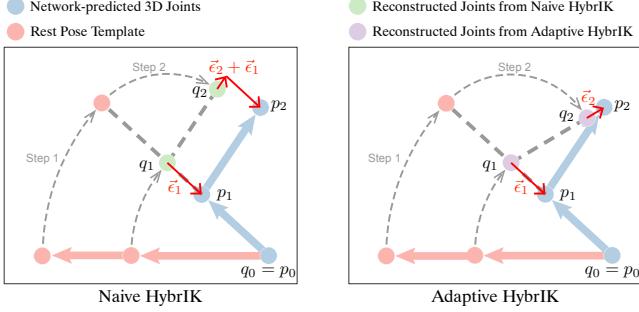
$$R_{pa(k)}^{-1}(p_k - p_{pa(k)}) = R_{pa(k),k}(t_k - t_{pa(k)}). \quad (11)$$

Let  $\vec{p}_k = R_{pa(k)}^{-1}(p_k - p_{pa(k)})$  and  $\vec{t}_k = (t_k - t_{pa(k)})$ , we can solve the relative rotation via Eq. 6:

$$R_{pa(k),k} = \mathcal{D}(\vec{p}_k, \vec{t}_k, \phi_k), \quad (12)$$

where  $\phi_k$  is the network-predicted *twist* angle for the  $k$ -th joint. The set of *twist* angle is denoted as  $\Phi = \{\phi_k\}_{k=1}^K$ . Since the rotation matrices are orthogonal, their inverse equals to their transpose, i.e.,  $R_{pa(k)}^{-1} = R_{pa(k)}^T$ , which keeps the solving process differentiable.

The whole process is named Naive HybrIK and summarized in Alg. 1. Note that we solve the relative rotation  $R_{pa(k),k}$  instead of the global rotation  $R_k$ . The reason is that if we directly decompose the global rotation, the resulting *twist* angle will depend on all ancestors' rotations along the



**Fig. 4. Example of the reconstruction error.** The joints in the rest pose are rotated to  $q_1$  and  $q_2$  by two steps. In the first step, due to the bone-length inconsistency, the reconstruction error is  $\vec{\epsilon}_1$ . In the second step, Naive HybrIK takes  $p_2 - p_1$  as the target direction, resulting in the accumulation of error  $\vec{\epsilon}_1 + \vec{\epsilon}_2$ . Instead, Adaptive HybrIK selects the reconstructed joint  $q_1$  to form the target direction  $p_2 - q_1$ , which reduces the error to only  $\vec{\epsilon}_2$ .

kinematic tree, which increases the variation of the distal limb joints and makes it difficult for the network to learn.

**Adaptive HybrIK.** Although the Naive HybrIK process seems effective, it follows an unstated hypothesis:  $\|p_k - p_{\text{pa}(k)}\| = \|t_k - t_{\text{pa}(k)}\|$ . Otherwise, there is no solution for Eq. 5. Unfortunately, in our case, the body-parts predicted by the 3D keypoint estimation method are not always consistent with the rest pose template. In Naive HybrIK, Eq. 6 can still be solved because the condition is turned into:

$$p_k - p_{\text{pa}(k)} = R_k(t_k - t_{\text{pa}(k)}) + \vec{\epsilon}_k, \quad (13)$$

where  $\vec{\epsilon}_k$  denotes the error in the  $k$ -th step, which has the same direction of  $p_k - p_{\text{pa}(k)}$  and  $\|\vec{\epsilon}_k\| = \||p_k - p_{\text{pa}(k)}\| - \|t_k - t_{\text{pa}(k)}\||$ . To analyze the reconstruction error, we compare the difference between the input pose  $\mathbf{P}$  and the reconstructed pose  $\mathbf{Q}$ :

$$\|\mathbf{P} - \mathbf{Q}\| \Leftrightarrow \sum_{k=1}^K \|p_k - q_k\|, \quad (14)$$

where  $\mathbf{Q} = FK(\mathbf{R}, \mathbf{T}) = FK(IK(\mathbf{P}, \mathbf{T}), \mathbf{T})$ . Combining Eq. 2 and Eq. 13, we have:

$$\begin{aligned} p_k - q_k &= p_{\text{pa}(k)} - q_{\text{pa}(k)} + \vec{\epsilon}_k \\ &= p_{\text{pa}^2(k)} - q_{\text{pa}^2(k)} + \vec{\epsilon}_{\text{pa}(k)} + \vec{\epsilon}_k \\ &= \dots = \sum_{i \in \mathcal{A}(k)} \vec{\epsilon}_i, \end{aligned} \quad (15)$$

where  $\text{pa}^2(k)$  denotes the parent index of the  $\text{pa}(k)$ -th joint, and  $\mathcal{A}(k)$  denotes the set of ancestors of the  $k$ -th joint. That means the difference between the input joint  $p_k$  and the reconstructed joint  $q_k$  will accumulate along the kinematic tree, which brings more uncertainty to the distal joint.

To address this error accumulation problem, we further propose Adaptive HybrIK. In Adaptive HybrIK, the target vector is adaptively updated with the newly reconstructed parent joints. Let  $\vec{p}_k = R_{\text{pa}(k)}^{-1}(p_k - q_{\text{pa}(k)})$  and  $\vec{t}_k$  the same as the one in the naive solution. In this way, the condition in Adaptive HybrIK can be formulated as:

$$p_k - q_{\text{pa}(k)} = R_k(t_k - t_{\text{pa}(k)}) + \vec{\epsilon}_k. \quad (16)$$

## Algorithm 2: Adaptive HybrIK

**Input:**  $\mathbf{P}, \mathbf{T}, \Phi$

**Output:**  $\mathbf{R}$

```

1 Determine  $R_0$ ;
2 for  $k$  along the kinematic tree do
3    $q_{\text{pa}(k)} \leftarrow R_{\text{pa}(k)}(t_{\text{pa}(k)} - t_{\text{pa}^2(k)}) + q_{\text{pa}^2(k)}$  ;
4    $\vec{p}_k \leftarrow R_{\text{pa}(k)}^{-1}(p_k - q_{\text{pa}(k)})$ ;
5    $\vec{t}_k \leftarrow (t_k - t_{\text{pa}(k)})$ ;
6    $R_{\text{pa}(k), k}^{sw} \leftarrow D^{sw}(\vec{p}_k, \vec{t}_k)$ ;
7    $R_{\text{pa}(k), k}^{tw} \leftarrow D^{tw}(\vec{t}_k, \phi_k)$ ;
8    $R_{\text{pa}(k), k} \leftarrow R_{\text{pa}(k), k}^{sw} R_{\text{pa}(k), k}^{tw}$ ;

```

Therefore, we have:

$$\begin{aligned} p_k - q_{\text{pa}(k)} &= q_k - q_{\text{pa}(k)} + \vec{\epsilon}_k \\ \Rightarrow p_k - q_k &= \vec{\epsilon}_k. \end{aligned} \quad (17)$$

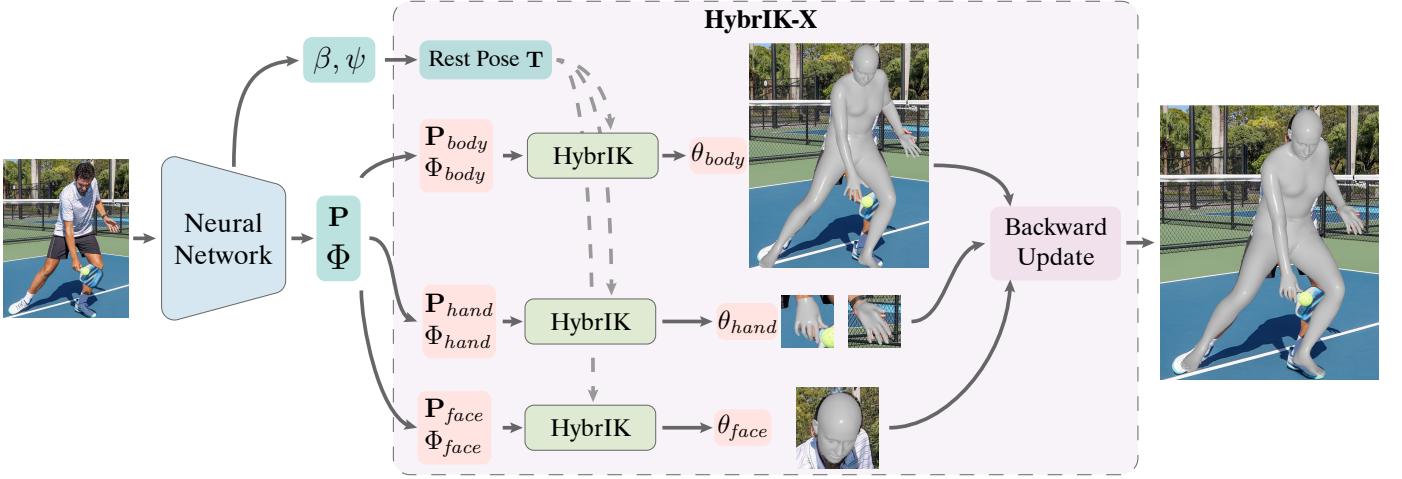
Compared to the naive solution (Eq. 15), the reconstructed error of the adaptive solution only depends on the current joint position. As illustrated in Fig. 4, in Naive HybrIK, once the parent joint is out of position, its children will continue this mistake. Instead, in Adaptive HybrIK, the solved relative rotation is always pointing towards the target joint and tries to reduce the error. We conduct empirical experiments in §4.5 to validate its robustness. The whole process of Adaptive HybrIK is summarized in Alg. 2.

### 3.2.3 Whole-body Inverse Kinematics

Adaptive HybrIK is accurate for body-only mesh recovery. It reduces the accumulated errors by adaptively updating the parent joints in a feedforward process. Nevertheless, extending it to whole-body mesh recovery is nontrivial. Although adaptive HybrIK tries to minimize the error in each step, the error won't be totally eliminated since we cannot fix the erroneous positions of the ancestor joints. As long as the bone length calculated from keypoints is different from the SMPL-X model, misalignment is inevitable. This misalignment is particularly pronounced when considering fine-grained face and hand mesh recovery since the kinematics tree is much deeper and the distal joints (e.g., fingers and face) are further away from the root joint.

**HybrIK-X.** To achieve well-aligned whole-body mesh recovery, we further present HybrIK-X. As depicted in Fig. 5, the core of HybrIK-X is a divide-and-conquer IK process. Specifically, we first divide the whole-body kinematic tree into four sub-trees (namely the left/right hands, face, and body), which have shorter lengths compared to the whole-body tree. By doing so, the distal joints in each sub-tree are closer to their corresponding root joint. Thus the sub-trees are more robust to noisy bone lengths and have a better model-image alignment. Subsequently, we apply HybrIK independently in each sub-tree. The recovered mesh of each sub-tree will align with its corresponding root joint (left/right wrists, head, and pelvis). Finally, we utilize a novel backward-updated algorithm to merge the results of all sub-trees and resolve the conflict joints.

The conflict joints refer to those joints that appear in two sub-trees concurrently, such as the left/right wrists and



**Fig. 5. Illustration of the whole-body reconstruction pipeline.** The whole-body 3D joints  $\mathbf{P}$ , twist angles  $\Phi$ , shape parameters  $\beta$ , and expression parameters  $\psi$  are regressed from a holistic model. These results are split to three kinematic sub-trees and employ HybrIK independently. We solve the conflict joint positions and merge the solved poses from each sub-tree with the proposed backward update technique.

head. These joints are considered both the root joints of the sub-trees and the distal joints of the body-tree. If we directly combine the solved rotations of 4 sub-trees, the positions of the distal joints will be only determined by the results from the body-tree because the blend skinning function of SMPL-X depends on the feedforward FK process. In order to preserve the well-aligned results from the sub-trees, we propose a backward-updated algorithm for recalculating the rotations of the parents of the conflict joints.

Given the estimated position of the distal joint  $p_k$  and the reconstructed position of its grandparent joint  $q_{pa^2(k)}$ , our objective is to recalculate the rotation  $R_{pa(k)}^*$  that satisfies:

$$R_{pa(k)}^* = \underset{R_{pa(k)}}{\operatorname{argmin}} \|R_{pa(k)}\vec{t}_{pa(k)} + q_{pa^2(k)} - p_{pa(k)}\|^2, \quad (18)$$

$$\text{s.t. } \exists R_k \in \mathcal{SO}(3), p_k = R_k \vec{t}_k + R_{pa(k)}^* \vec{t}_{pa(k)} + q_{pa^2(k)}.$$

By recalculating the rotation of the parent joint, the distal joint position will be consistent with the root position of the sub-tree. However, Eq. 18 involves the  $\mathcal{SO}(3)$  space and is not easy to solve. To make this problem solvable in a differentiable manner, we reformulate Eq. 18 to an equivalent problem:

$$q_{pa(k)}^* = \underset{q_{pa(k)}}{\operatorname{argmin}} \|q_{pa(k)} - p_{pa(k)}\|^2, \quad (19)$$

$$\text{s.t. } \begin{cases} \|q_{pa(k)}^* - q_{pa^2(k)}\| = \|\vec{t}_{pa(k)}\|, \\ \|p_k - q_{pa(k)}^*\| = \|\vec{t}_k\|. \end{cases}$$

The analytical solution of the position of the parent joint  $q_{pa(k)}^*$  can be easily solved and computed differentiably. Detailed derivations are provided in appendix §B. Once we obtain  $q_{pa(k)}^*$ , we follow the proposed twist-and-swing decomposition to calculate  $R_{pa(k)}$  and  $R_k$ .

**Robust Jaw Pose Estimation.** The jaw rotation in the SMPL-X model relies on two joints, namely jaw and head. However, accurately determining the position of the head joint can be challenging, particularly when people are wearing hats or looking upwards. Therefore, using Eq. 7 and Eq. 8

to calculate the swing rotation can result in an anomalous mouth-opening posture. To address this issue, we calculate jaw rotation by measuring the angle of mouth opening. In particular, we replace the jaw node in the kinematic tree with two nodes picked from the SMPL-X mesh (mouth top and mouth bottom). We then utilize these two joints to determine the swing axis and swing angle, and subsequently obtain the swing rotation for the jaw joint using Eq. 9.

### 3.3 Learning Framework

The overall framework of our approach is illustrated in Fig. 2. Firstly, a neural network is utilized to predict 2.5D joints  $\mathbf{P}_{2.5D}$ , twist angles  $\Phi$ , shape parameters  $\beta$ , expression parameters  $\psi$ , and the initial camera parameter  $s^0$ . The 2.5D joints  $\mathbf{P}_{2.5D}$  and the initial camera parameter  $s^0$  are sent to the iterative camera estimation module to obtain the final camera prediction and the 3D joints  $\mathbf{P}$ . Secondly, the shape and expression parameters are used to obtain the rest pose  $T$  from the SMPL/SMPLEX model. Then, by combining  $\mathbf{P}$ ,  $T$  and  $\Phi$ , we employ HybrIK/HybrIK-X to solve the rotations  $\mathbf{R}$  of the 3D body, i.e., the pose parameters  $\theta$ . Finally, with the function  $\mathcal{M}(\theta, \beta, \psi)$  provided by the SMPL/SMPLEX model, the whole-body mesh  $M$  is obtained. The reconstructed pose  $Q$  can be obtained from  $M$  by FK or a regressor, which is guaranteed to be realistic. Since HybrIK and HybrIK-X are differentiable, the whole framework is trained in an end-to-end manner.

**Regression-based 3D Keypoint Estimation.** Previous approaches to 3D keypoint estimation use heatmaps to represent the likelihood of joint positions. However, this heatmap representation costs a significant computational burden. Besides, it limits the output range within the input bounding box, which fails in the truncated scenarios where part of the human body is outside the input image. In real-world challenging scenarios, object detection methods are not perfect and always generate truncated bounding boxes when people are occluded or partially visible.

To reduce the computational cost and improve the robustness to real-world challenging scenarios, we adopt RLE [110], a simple yet effective regression paradigm. The conventional RLE leverages a fully-connected layer to estimate the joint coordinates  $p_k \in \mathbb{R}^3$  and the standard deviation of each coordinate value  $\sigma_k \in \mathbb{R}^3$ . However, due to the limited diversity of current 3D human pose datasets, predicting three-dimensional deviations makes the learned distribution overfit to the training set, leading to an unstable training process. Here, we propose to use one-dimensional deviations to reduce the parameters of the learnable distribution and prevent overfitting, i.e.,  $\sigma_k \in \mathbb{R}$ . When calculating the negative log-likelihood loss, we assume the coordinates of the three axes share the same standard deviation  $\sigma_k$ . The loss to train the 3D keypoint is formulated as:

$$\mathcal{L}_{pose} = -\sum_{k=1}^K \log Q(\bar{p}_k) - \log G_\psi(\bar{p}_k) + 3 \cdot \log \sigma_k, \quad (20)$$

where  $Q(\cdot)$  is the probability density function of the standard Laplace distribution,  $G_\psi$  is the distribution learned by RLE [110], and  $\bar{p}_k = (p_k - \hat{p}_k)/\sigma_k$  and  $\hat{p}_k$  denotes the ground-truth joint position. In practice, the network output for each joint is a 2.5D coordinate, i.e., the two-dimensional pixel coordinate with a relative depth value. To obtain the 3D coordinates, we back-project the 2.5D point to the 3D space with the estimated camera parameters.

With the regression paradigm, we get rid of the heatmap representation, and the model can be trained to infer the invisible body joints in challenging scenarios. Besides, the predicted deviation  $\sigma_k$  can serve as the uncertainty index to evaluate the reliability of predicted keypoints, which is essential for downstream applications.

**Iterative Camera Estimation.** To back-project the 2.5D points onto the 3D space, we employ the weak-perspective camera model with a focal length of 1 m. In contrast to previous work [6], [7], [10], we solely regress the scale factor  $s \in \mathbb{R}$  and discard the translation, since we can determine the 2D position of the person by the root joint position. However, inferring the scale factor  $s$  from a monocular RGB image is an ill-posed problem. To address this, we propose an iterative camera estimation method (ICE) that fully exploits the 2D visual cues, human body structure, and the power of the deep neural network for accurate and stable camera estimation. Specifically, given the currently estimated scale factor  $s^t$  in the  $t$ -th step, we back-project the 2.5D points to 3D points:

$$\mathbf{P}^t = \{p_k^t\}_{k=1}^K = \Pi^{-1}(\mathbf{P}_{2.5D}; s^t), \quad (21)$$

where  $\Pi(\cdot)$  is the projection function and  $\Pi^{-1}(\cdot)$  denotes the back-projection function. Since the initial scale prediction is not correct, the projected 3D joints  $\mathbf{P}^t$  might be wrong, e.g., overly small or overly large. We then input  $\mathbf{P}^t$  to HybIK-X to reconstruct the whole-body pose and retrieve reconstructed keypoints  $\mathbf{Q}^t$ , which satisfy the constraints imposed by the human body structure. We then use the least square method to analytically calculate the updated

$s^{t+1}$  that minimizes projection error between  $\mathbf{Q}^t$  and the detected 2D keypoints:

$$s^{t+1} = \arg \min_{s^*} \|\Pi(\Pi^{-1}(\mathbf{P}_{2.5D}; s^t); s^*) - \mathbf{P}_{2D}\|^2. \quad (22)$$

In this way, the human body structure and the 2D visual cues can provide a more accurate scale estimation. In the next iteration, a more accurate scale  $s^{t+1}$  can obtain more accurate back-projected keypoints  $\mathbf{P}^{t+1}$  and reconstructed keypoints  $\mathbf{Q}^{t+1}$ . Such an iterative update can alleviate the ambiguity in camera parameter estimation and generate stable results. The initially estimated  $s^0$  is regressed by the neural network and supervised by the  $\ell_2$  loss:

$$\mathcal{L}_{cam} = \|s^0 - \hat{s}\|^2, \quad (23)$$

where  $\hat{s}$  denotes the ground-truth scale factor.

**Twist Angle Estimation.** Instead of directly regressing the scalar value  $\phi_k$ , we choose to learn a 2-dimensional vector  $(\cos \phi_k, \sin \phi_k)$  to avoid the discontinuity problem. The  $\ell_2$  loss is applied:

$$\mathcal{L}_{tw} = \frac{1}{K} \sum_{k=1}^K \|(\cos \phi_k, \sin \phi_k) - (\cos \hat{\phi}_k, \sin \hat{\phi}_k)\|^2, \quad (24)$$

where  $\hat{\phi}_k$  denotes the ground-truth twist angle for the  $k$ -th joint.

**Collaboration with SMPL-X.** The SMPL-X model allows us to obtain the rest pose skeleton with the additive offsets according to the shape parameters  $\beta$  and expression parameters  $\psi$ :

$$\mathbf{T} = W(\bar{M}_{\mathbf{T}} + B_S(\beta) + B_E(\psi)), \quad (25)$$

where  $\bar{M}_{\mathbf{T}}$  is the mesh vertices of mean rest pose,  $B_S(\beta)$  and  $B_E(\psi)$  are the blend shape functions provided by SMPL-X. Then the pose parameters  $\theta$  are calculated by HybIK-X in a differentiable manner. In the training phase, we supervise the shape parameters  $\beta$ :

$$\mathcal{L}_{shape} = \|\beta - \hat{\beta}\|^2, \quad (26)$$

the expression parameters  $\psi$ :

$$\mathcal{L}_{exp} = \|\psi - \hat{\psi}\|^2, \quad (27)$$

and the rotation parameters  $\theta$ :

$$\mathcal{L}_{rot} = \|\theta - \hat{\theta}\|^2. \quad (28)$$

The overall loss of the learning framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{pose} + \mu_1 \mathcal{L}_{cam} + \mu_2 \mathcal{L}_{shape} + \mu_3 \mathcal{L}_{exp} + \mu_4 \mathcal{L}_{rot} + \mu_5 \mathcal{L}_{tw}, \quad (29)$$

where  $\mu_1, \mu_2, \mu_3, \mu_4$  and  $\mu_5$  are weights of the loss items.

### 3.4 Implementation Details

Here we elaborate more implementation details. We use HRNet-W48 [111] as the network backbone by default, initialized with ImageNet pre-trained weights. The HRNet output is fed to an average pooling layer, followed by the fully-connected layers to regress  $\beta, \psi, \phi, \mathbf{P}_{2.5D}$  and  $s^0$ . The input image is resized to  $256 \times 256$ . The learning rate is set to  $1 \times 10^{-3}$  at first and reduced by a factor of 10 at the 90th and 120th epoch. We use the Adam solver and train for 140 epochs, with a mini-batch size of 32 per GPU and 4 GPUs in total. In all experiments,  $\mu_1 = \mu_2 = \mu_3 = 1$  and  $\mu_4 = \mu_5 = 1 \times 10^{-2}$ . Implementation is in PyTorch.

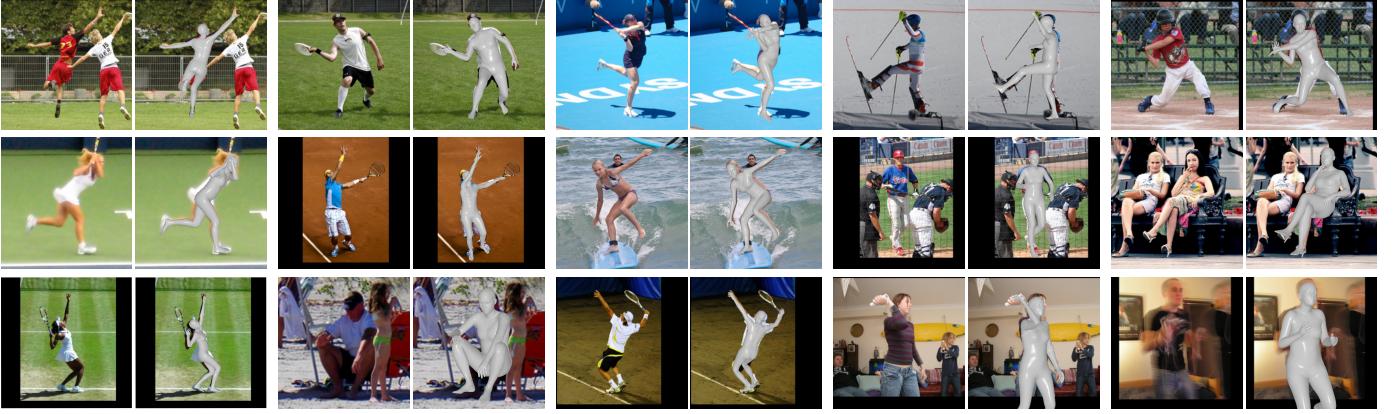


Fig. 6. Qualitative results of body-only mesh recovery on the MSCOCO validation set with challenging poses, occlusions, truncations, and motion blurs.

TABLE 1

Benchmark of state-of-the-art models on 3DPW, Human3.6M and MPI-INF-3DHP datasets. “\*” denotes the method is trained on different datasets. “-” shows the results that are not available.

Method	3DPW			Human3.6M			MPI-INF-3DHP		
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	PA-MPJPE ↓	MPJPE ↓	PCK ↑	AUC ↑	MPJPE ↓	
SMPLify [5]	-	-	-	82.3	-	-	-	-	
HMR [6]	81.3	130.0	-	56.8	88.0	72.9	36.5	124.2	
Pavlakos et al. [53]	-	-	-	75.9	-	-	-	-	
SPIN [7]	59.2	96.9	116.4	41.1	-	76.4	37.1	105.2	
I2L-MeshNet [112]*	58.6	93.2	-	41.7	55.7	-	-	-	
KAMA [57]	51.1	-	97.0	40.2	-	-	-	-	
ROMP [56]	47.3	76.7	93.4	-	-	-	-	-	
METRO [58]	47.9	77.1	88.2	36.7	54.0	-	-	-	
PARE [9]	46.5	74.5	88.6	-	-	-	-	-	
HybrIK (ResNet-34)	44.6	72.5	89.1	33.7	55.3	86.5	46.9	93.3	
HybrIK (HRNet-W48)	<b>41.8</b>	<b>71.6</b>	<b>82.3</b>	<b>29.8</b>	<b>47.0</b>	<b>87.1</b>	<b>47.3</b>	<b>91.0</b>	

TABLE 2

Quantitative comparisons with state-of-the-art body-only methods on the AGORA dataset.

Method	NMVE ↓	NMJE ↓	MVE ↓	MPJPE ↓
HMR [6]	217.0	226.0	173.6	180.5
SPIN [7]	193.4	199.2	148.9	153.4
EFT [113]	196.3	203.6	159.0	165.4
PARE [9]	167.7	174.0	140.9	146.2
SPEC [114]	126.8	133.7	106.5	112.3
ROMP [56]	113.6	118.8	103.4	108.1
PyMAF [55]	200.2	207.4	168.2	174.2
BEV [115]	108.3	113.2	100.7	105.3
Hand4Whole [87]	90.2	95.5	84.8	89.8
CLIFF [116]	83.5	89.0	76.0	81.0
HybrIK	<b>81.2</b>	<b>84.6</b>	<b>73.9</b>	<b>77.0</b>

## 4 EMPIRICAL EVALUATION

In this section, we first describe the datasets employed for training and quantitative evaluation. Next, we compare HybrIK and HybrIK-X with state-of-the-art approaches on body-only, hand-only, and whole-body mesh recovery benchmarks. Finally, ablation experiments are conducted to evaluate HybrIK and HybrIK-X.

### 4.1 Datasets

Following previous work, we train the body-only HybrIK on the Human3.6M [17], 3DPW [16], MPI-INF-3DHP [18], MSCOCO [117], and AGORA [21] datasets. For hand-only HybrIK, we train and evaluate on FreiHAND [19] and HO3D-v2 [20] datasets. For whole-body HybrIK-X, we use Human3.6M [17], 3DPW [16], MPI-INF-3DHP [18], MSCOCO [117], and AGORA [21] datasets for training. Detailed descriptions of the datasets are provided appendix §C.

### 4.2 Evaluation on Body-only Mesh Recovery

To make a fair comparison with previous body-only mesh recovery methods, we use a regressor to obtain the 14 LSP joints from the body mesh for the evaluation on 3DPW and Human3.6M datasets and 17 joints for the MPI-INF-3DHP dataset. Procrustes aligned mean per joint position error (PA-MPJPE), mean per joint position error (MPJPE), percentage of correct keypoints (PCK), and area under curve (AUC) are reported to evaluate the 3D pose results. Per/Mean vertex error (PVE/MVE) is reported to evaluate the entire estimated body mesh. We further conduct experiments on the official AGORA test set. Normalized mean joint error (NMJE) and normalized mean vertex error (NMVE) are additionally reported.

TABLE 3  
Quantitative comparisons with state-of-the-art whole-body methods on the AGORA dataset.

Method	NMVE ↓		NMJE ↓		MVE ↓		MPJPE ↓					
	FB	B	FB	B	FB	B	F	LH/RH	FB	B	F	LH/RH
SMPLify-X [3]	333.1	263.3	326.5	256.5	236.5	187.0	48.9	48.3/51.4	231.8	182.1	52.9	46.5/49.6
ExPose [84]	265.0	184.8	263.3	183.4	217.3	151.5	51.1	74.9/71.3	215.9	150.4	55.2	72.5/68.8
FrankMocap [14]	-	207.8	-	204.0	-	168.3	-	54.7/55.7	-	165.2	-	52.3/53.1
PIXIE [10]	233.9	173.4	230.9	171.1	191.8	142.2	50.2	49.5/49.0	189.3	140.3	54.5	46.4/46.0
Hand4Whole [87]	144.1	96.0	141.1	92.7	135.5	90.2	41.6	46.3/48.1	132.6	87.1	46.1	44.3/46.2
PyMAF-X [15]	141.2	94.4	140.0	93.5	125.7	84.0	35.0	44.6/45.6	124.6	83.2	37.9	42.5/43.7
HybrIK-X	<b>120.5</b>	<b>73.7</b>	<b>115.7</b>	<b>72.3</b>	<b>112.1</b>	<b>68.5</b>	<b>37.0</b>	<b>46.7/47.0</b>	<b>107.6</b>	<b>67.2</b>	<b>38.5</b>	<b>41.2/41.4</b>

TABLE 4  
Quantitative comparisons with state-of-the-art methods on the FreiHAND dataset.

Method	PA-PVE ↓	PA-MPJPE ↓	F-Score @5mm ↑
Hand-only	FreiHAND [19]	10.7	-
	Hasson et al. [71]	13.2	-
	Boukhayma et al. [70]	13.0	-
	Pose2Mesh [118]	7.8	0.674
	I2L-MeshNet [112]	7.6	0.681
Whole-body	METRO [58]	6.3	0.731
	ExPose [84]	11.8	0.484
	Zhou et al. [85]	-	15.7
	FrankMocap [14]	11.6	0.553
	PIXIE [10]	12.1	0.468
	Hand4Whole [87]	7.7	0.664
HybrIK	PyMAF [15]	8.1	0.638
	HybrIK	<b>6.2</b>	<b>6.0</b>
			<b>0.761</b>

In Tab. 1, we compare our method with previous 3D human pose and shape estimation methods, including both model-based and model-free methods, on 3DPW, Human3.6M, and MPI-INF-3DHP datasets. Without bells and whistles, our method surpasses all previous state-of-the-art methods by a large margin on all three datasets. It is worth noting that our method improves 4.7 mm PA-MPJPE (**10.1%** relative improvement) on the 3DPW dataset, which shows that it is accurate and reliable to recover body mesh through inverse kinematics.

In Tab. 2, we further compare our method on the SMPL track of the AGORA test set. Compared to other 3D pose datasets, AGORA contains more challenging scenarios with severe occlusions and truncations. HybrIK shows consistent improvements on this dataset. Qualitative results are shown in Fig. 6.

#### 4.3 Evaluation on Hand-only Mesh Recovery

To validate the generalization of our inverse kinematics solution, we conduct experiments on two widely-used hand pose benchmarks: FreiHAND [19] and HO3D [20]. MPJPE and PVE are reported for evaluation. F-Score [123] is also reported to evaluate the harmonic mean between the predicted vertices and the ground-truth vertices.

The comparisons of HybrIK with previous state-of-the-art methods are reported in Tab. 4 and Tab. 5. It is noteworthy that, as observed in prior research [15], [112], recent hand-only methods [58], [112] that adopt non-parametric representation exhibit a numerical advantage

TABLE 5  
Quantitative comparisons with state-of-the-art methods on the HO-3D dataset.

Method	PA-PVE ↓	PA-MPJPE ↓	F-Score @5mm ↑
Pose2Mesh [118]	1.27	1.25	0.441
I2L-MeshNet [112]	1.39	1.12	0.409
I2UV-HandNet [119]	1.01	<b>0.99</b>	0.500
METRO [58]	1.11	1.04	0.484
Hampali et al. [20]	1.06	1.07	0.506
Hasson et al. [71]	1.12	1.10	0.464
Hasson et al. [120]	1.14	1.14	0.428
ArtiBoost [121]	1.09	1.14	0.488
Keypoint Trans. [122]	-	1.08	-
HybrIK	<b>0.96</b>	<b>0.99</b>	<b>0.550</b>

over parametric-based methods. HybrIK significantly outperforms the most accurate whole-body method by 1.5 mm MPJPE (**19.5%** relative improvement) on the FreiHAND dataset. Additionally, even when compared against the state-of-the-art hand-only methods, HybrIK exhibits a 7.5% relative improvement. Compared to the methods designed specifically for challenging interaction scenarios on the HO3D dataset, HybrIK also exhibits state-of-the-art performance.

#### 4.4 Evaluation on Whole-body Mesh Recovery

To evaluate our hybrid inverse kinematics solution on whole-body mesh recovery, we conduct experiments on the SMPL-X track of the AGORA test set. The evaluation on AGORA is affected by the detection results since there are multiple persons on one image. We follow previous work [15], [87] and use the same detection results for a fair comparison.

The evaluation is depicted in Tab. 3. Note that previous methods use separate expert networks to handle body, face, and hand estimation. They cost higher computational resources and longer running time. HybrIK-X is a one-stage approach and demonstrates a significant superiority over the state-of-the-art whole-body methods, achieving **24.3** mm and **20.7** mm improvement in full-body NMJE and NVME, respectively. For the fine-grained face and hand results, HybrIK-X obtains comparable performance against the most accurate method with a much smaller input resolution and less computation. Qualitative results on the MSCOCO validation set are shown in Fig. 7. Qualitative comparison



Fig. 7. Qualitative results of whole-body mesh recovery on the MSCOCO validation set.

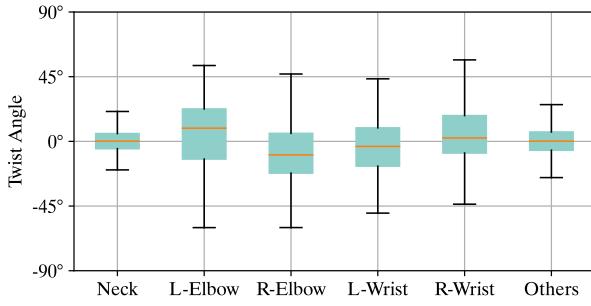


Fig. 8. Distribution of twist angles. Only a few joints have a range over  $30^\circ$ . Other joints have a limited range of twist angle.

with state-of-the-art approaches are provided in appendix §E. Detailed comparisons of computation complexity are reported in §4.5.

#### 4.5 Ablation Study

In this study, we evaluate the effectiveness of the twist-and-swing decomposition and the proposed inverse kinematics algorithms. Evaluation is conducted on the AGORA validation set by default as it contains challenging in-the-wild scenarios. More experimental results are provided in appendix §D.

**Analysis of the twist rotation.** To demonstrate the effectiveness of twist-and-swing decomposition, we first count the distribution of the *twist* angle in the AGORA validation set. The distribution is illustrated in Fig. 8. As expected, due to the physical limitation, only neck, elbow and wrist have a wide range of variations. All other joints have a limited range of *twist* angle (less than  $30^\circ$ ). It indicates that the *twist* angle can be reliably estimated.

TABLE 6  
Reconstruction error with different *twist* angle. The accurate *twist* angles significantly reduce the reconstruction error.

	Random Twist			Estimated Twist			Zero Twist		
	24 jts	14 jts	Vert.	24 jts	14 jts	Vert.	24 jts	14 jts	Vert.
Error	0.1	40.0	67.3	0.1	6.1	10.0	0.1	6.8	12.1

TABLE 7  
Robustness to noisy joints. Mean errors of body and hand joints are reported.

	Body / Hand MPJPE			
	GT Joints	$\pm 10$ mm	$\pm 20$ mm	$\pm 30$ mm
Naive HybrIK	0.1/0.1	44.4/57.6	74.5/89.5	104.7/121.9
Adaptive HybrIK	0.1/0.1	21.6/16.5	40.6/32.3	59.0/47.8
HybrIK-X	0.1/0.1	21.3/13.6	39.6/26.9	57.6/40.4

Besides, we develop an experiment to see how the *twist* angles affect the reconstructed pose and shape. We take the ground-truth 24 SMPL joints and shape parameters as the input of the HybrIK process. As for the *twist* angle, we compare random values in  $[-\pi, \pi]$  and the values estimated by the network. We evaluation the mean error of the reconstructed 24 SMPL joints, the 14 LSP joints, the body mesh and the *twist* angle. Here, following previous work [5], [6], [7], the 14 LSP joints are regressed from the body mesh by a pretrained regressor. Quantitative results are reported in Tab. 6. It shows that the regressed *twist* angles significantly reduce the error on the mesh vertices and the LSP joints that regressed from the mesh. Since most of the twist angles are close to zeros, the zero twist angles produce acceptable performance. Notice that the wrong *twist* angles do not affect the reconstructed SMPL joints. Only the *swing* rotations change the joint locations.

**Robustness to Noisy Joint Positions.** To demonstrate the superiority of HybrIK-X over Adaptive and Naive HybrIK, we compare the reconstruction errors with inputs in different noise levels on the AGORA validation set. We add jitters to the ground-truth joint positions and then feed them to the IK algorithms. Quantitative comparisons on whole-body recovery are reported in Tab. 7. We report the reconstruction errors of body and hand joints separately. It shows that when the input joints are correct, all three algorithms introduce negligible errors. As the noise level increases, HybrIK-X and Adaptive HybrIK are more robust than Naive HybrIK. For body reconstruction, HybrIK-X has similar performance to Adaptive HybrIK. For fine-grained hand reconstruction, HybrIK-X exhibits better robustness against the noisy joint positions.

**Error correction capability.** In this experiment, we examine the error correction capability of the HybrIK algorithm. The HybrIK algorithm is fed with the 3D joints, *twist* angles and shape parameters that predicted by the neural network. Additionally, we apply the SMPLify [5] algorithm on the

TABLE 8  
The error correction capability.

	Predicted Pose	HybrIK	SMPLify [5]
MPJPE (24 joints) ↓	73.1 mm	<b>66.3 mm</b>	100.1 mm

TABLE 9

**Ablation experiments on the truncated AGORA validation set.**  
“RLE<sup>-</sup>” denotes our improved version of RLE with reduced distribution parameters.

Method	Validation w. Trunc.		Test	
	MVE ↓	MPJPE ↓	MVE ↓	MPJPE ↓
HybrIK-X (Heatmap-based)	100.4	99.8	134.1	127.5
HybrIK-X (RLE)	91.9	82.7	113.9	109.7
HybrIK-X (RLE <sup>-</sup> )	<b>89.9</b>	<b>80.4</b>	<b>112.1</b>	<b>107.6</b>

predicted pose and compare it to our method. As shown in Tab. 8, the error of reconstructed joints after HybrIK is reduced the error to 66.3mm, while SMPLify raises the error to 100.1 mm. The error correction capability of HybrIK comes from the fact that the network may predict unrealistic body pose, e.g., left-right asymmetry and abnormal limbs proportions. In contrast, the rest pose is generated by the parametric statistical body model, which guarantees that the reconstructed pose is consistent with the realistic body shape distribution. Since our proposed framework is agnostic to the way we obtain 3D joints, we can improve the performance of any 3D keypoint estimation approaches.

**Effectiveness of the Regression Paradigm.** We further evaluate the effectiveness of the regression paradigm. Since the AGORA test set does not provide ground-truth bounding boxes, the model performance will be affected by the object detection results. Besides, the test set is quite challenging because the persons in it are severely occluded and truncated. Therefore, the AGORA test set can serve as the benchmark to evaluate the robustness of the mesh recovery algorithms. Moreover, we augment the current validation set with challenging truncated bounding boxes. We simulate truncations by cropping each frame with a truncated window. The areas of the cropped windows are 3/4 of the original ones. Quantitative comparisons about different 3D keypoint estimation paradigms are reported in Tab. 9. Heatmap cannot represent the joints outside the input bounding box. Therefore, there is a significant performance degradation when using heatmaps to obtain joint positions. Additionally, the improved version of RLE shows better performance than the original RLE since it introduces fewer parameters to the distribution and avoids overfitting. It is demonstrated that our method is robust to challenging applications with occlusions and truncations. Qualitative comparisons between the heatmap-based paradigm and the regression-based paradigm are provided in appendix §E.

**Effectiveness of Iterative Camera Estimation.** To study the effectiveness of the proposed iterative camera estimation (ICE), we evaluate the performance with different iterative steps. We calculate the mean error of the estimated distance

TABLE 10  
Results of depth estimation on the AGORA validation set.

	ICE Steps					
	0	1	2	3	4	5
Error ↓	183.91	181.10	179.51	178.60	178.18	178.09

TABLE 11  
**Computation complexity and model parameters.** Full-body mean vertex error (FB-MVE) is also listed.

Method	FLOPS ↓	#Params ↓	FB-MVE
PIXIE [10]	31.0G	192.9M	191.8
PyMAF-X [15]	46.5G	203.6M	125.7
HybrIK-X (One-Stage)	<b>22.7G</b>	<b>76.1M</b>	<b>112.1</b>

between the human body and the camera. Quantitative results are summarized in Tab. 10. “0 step” is equivalent to direct regression without the iterative update. We can observe that ICE can improve the accuracy of the camera distance estimation. In practice, 3 steps are accurate enough.

**Computation Complexity.** The experimental results of computation complexity and model parameters are listed in Tab. 11. The proposed one-stage HybrIK-X achieves more accurate whole-body mesh recovery results with significantly lower computation complexity and fewer model parameters. Specifically, the total FLOPs are reduced by 26.8%, and the parameters are reduced by 60.5%. The efficiency and effectiveness of HybrIK-X are of great value in downstream applications.

## 5 CONCLUSION

In this paper, we present a hybrid analytical-neural inverse kinematics framework, HybrIK, for body-only mesh recovery. This framework is further extended to whole-body mesh recovery and named HybrIK-X. HybrIK and HybrIK-X transform the 3D joint locations to a pixel-aligned accurate human body mesh via inverse kinematics, and then obtains a more accurate and realistic 3D skeleton from the reconstructed 3D mesh with forward kinematics, closing the loop between the 3D skeleton and the parametric body model. To evaluate the effectiveness of our approach, we perform experiments on various benchmarks for body-only, hand-only, and whole-body scenarios. The results indicate that our approach outperforms existing state-of-the-art approaches by a significant margin. Moreover, the proposed approach is fully differentiable and uses a one-stage network to recover the whole-body mesh, making it considerably more efficient than existing approaches. Overall, we believe our approach can serve as a strong baseline for future research and offers a new perspective on whole-body mesh recovery.

## APPENDIX A RIGID REGISTRATION OF GLOBAL ROTATION

In the SMPL/SMPL-X model [2], [3], the pose parameters  $\theta$  control the rotations of the rigid body parts. The three joints

named spine, left hip and right hip constitute a rigid body part, which is controlled by the global root rotation. Consequently, the global rotation can be ascertained by registering the rest pose template of spine, left hip and right hip to the predicted locations of these three joints. Let  $t_1$ ,  $t_2$  and  $t_3$  denote their locations in the rest pose template, and  $p_1$ ,  $p_2$  and  $p_3$  denote the predicted locations. Our objective is to identify a rigid rotation that optimally aligns the two sets of joints. Here, we assume the root joint of the predicted pose and the rest pose are aligned. Hence, the problem is formulated as:

$$R_0 = \arg \min_{R \in \mathbb{SO}^3} \sum_{i=1}^3 \|p_i - Rt_i\|_2^2. \quad (30)$$

This formula can be written in matrix form:

$$R_0 = \arg \min_{R \in \mathbb{SO}^3} \|P_0 - RT_0\|_F^2, \quad (31)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm,  $P_0$  denotes  $[p_0 \ p_1 \ p_2]$ , and  $T_0$  denotes  $[t_0 \ t_1 \ t_2]$ . Let us simplify the expression in Eq. 31 as:

$$\begin{aligned} & \min_{R \in \mathbb{SO}^3} \|P_0 - RT_0\|_F^2 \\ \Leftrightarrow & \min_{R \in \mathbb{SO}^3} \text{trace}((P_0 - RT_0)^T (P_0 - RT_0)) \\ \Leftrightarrow & \min_{R \in \mathbb{SO}^3} \text{trace}(P_0^T P_0 + T_0^T T_0 - 2P_0^T R T_0). \end{aligned} \quad (32)$$

Note that  $P_0^T P_0$  and  $T_0^T T_0$  are independent of  $R$ . Thus the original problem is equivalent to:

$$\begin{aligned} & \arg \min_{R \in \mathbb{SO}^3} \|P_0 - RT_0\|_F^2 \\ \Leftrightarrow & \arg \max_{R \in \mathbb{SO}^3} \text{trace}(P_0^T R T_0). \end{aligned} \quad (33)$$

Further, we can leverage the property of the matrix trace,

$$\text{trace}(P_0^T R T_0) = \text{trace}(R T_0 P_0^T). \quad (34)$$

Then, we apply Singular Value Decomposition (SVD) to the joint locations:

$$T_0 P_0^T = U \Lambda V^T. \quad (35)$$

The problem is equivalent to:

$$\begin{aligned} & \arg \max_{R \in \mathbb{SO}^3} \text{trace}(R T_0 P_0^T) \\ \Leftrightarrow & \arg \max_{R \in \mathbb{SO}^3} \text{trace}(R U \Lambda V^T) \\ \Leftrightarrow & \arg \max_{R \in \mathbb{SO}^3} \text{trace}(\Lambda V^T R U). \end{aligned} \quad (36)$$

Note that  $U$ ,  $V$  and  $R$  are orthogonal matrices, so  $M = V^T R U$  is also an orthogonal matrix. Then, for all  $1 \leq j \leq 3$  we have:

$$\begin{aligned} m_j^T m_j &= 1 = \sum_{i=1}^3 m_{ij}^2 \\ \Rightarrow m_{ij}^2 &\leq 1 \Rightarrow |m_{ij}| \leq 1. \end{aligned} \quad (37)$$

Besides,  $\Lambda$  is a diagonal matrix with non-negative values, i.e.,  $\lambda_1, \lambda_2, \lambda_3 \geq 0$ . Therefore:

$$\begin{aligned} \text{trace}(\Lambda V^T R U) &= \text{trace}(\Lambda M) \\ &= \sum_{i=1}^3 \lambda_i m_{ii} \leq \sum_{i=1}^3 \lambda_i. \end{aligned} \quad (38)$$

The trace is maximized if  $m_{ii} = 1, \forall 1 \leq i \leq 3$ . That means  $M = \mathcal{I}$ , where  $\mathcal{I}$  is the identity matrix. Finally, the optimal rotation  $R_0$  is:

$$\begin{aligned} V^T R_0 U &= \mathcal{I} \\ \Rightarrow R_0 &= V U^T. \end{aligned} \quad (39)$$

## APPENDIX B BACKWARD-UPDATED ALGORITHM

In the backward-updated algorithm of HybrIK-X, our aim is to analytically calculate the position of the parent joint, denoted as  $q_{\text{pa}(k)}^*$ . Recall that  $q_{\text{pa}(k)}^*$  should satisfy the following equation:

$$\begin{aligned} q_{\text{pa}(k)}^* &= \underset{q_{\text{pa}(k)}}{\text{argmin}} \|q_{\text{pa}(k)} - p_{\text{pa}(k)}\|^2, \\ \text{s.t. } & \begin{cases} \|q_{\text{pa}(k)}^* - q_{\text{pa}^2(k)}\| = \|\vec{t}_{\text{pa}(k)}\|, \\ \|p_k - q_{\text{pa}(k)}^*\| = \|\vec{t}_k\|. \end{cases} \end{aligned} \quad (40)$$

To simplify the notation, we define  $A = q_{\text{pa}^2(k)}$ ,  $B = p_{\text{pa}(k)}$ ,  $B^* = q_{\text{pa}(k)}^*$ , and  $C = p_k$  for the subsequent derivation. The aforementioned equation can be reformulated as:

$$B^* = \underset{q_{\text{pa}(k)}}{\text{argmin}} \|q_{\text{pa}(k)} - B\|^2, \quad (41)$$

$$\text{s.t. } \begin{cases} \|B^* - A\| = \|\vec{t}_{\text{pa}(k)}\|, \\ \|C - B^*\| = \|\vec{t}_k\|. \end{cases} \quad (42)$$

The vector  $\overrightarrow{AB^*}$  can be orthogonally decomposed with respect to  $\overrightarrow{AC}$  as follows:

$$\overrightarrow{AB^*} = \vec{v}_{\parallel} + \vec{v}_{\perp}, \quad (43)$$

where  $\vec{v}_{\parallel}$  is parallel to  $\overrightarrow{AC}$  and  $\vec{v}_{\perp}$  is perpendicular to  $\overrightarrow{AC}$ . We introduce a point  $D$ , such that  $\overrightarrow{AD} = \vec{v}_{\parallel}$  and  $\overrightarrow{DB^*} = \vec{v}_{\perp}$ . Since  $\vec{v}_{\parallel}$  is parallel to  $\overrightarrow{AC}$ , it can be expressed as  $\overrightarrow{AD} = m \overrightarrow{AC}$ , with  $m \in [0, 1]$  representing the corresponding scalar. Consequently, by determining  $k$  and  $\vec{v}_{\perp}$ , we can ascertain the position of  $B^*$ .

According to the norm constraints in Eq. 42, we can determine  $m$  as:

$$\begin{aligned} \|\vec{t}_{\text{pa}(k)}\|^2 - m^2 \overrightarrow{AC}^2 &= \|\vec{t}_k\|^2 - (1-m)^2 \overrightarrow{AC}^2, \\ \Rightarrow m &= \frac{\|\vec{t}_{\text{pa}(k)}\|^2 - \|\vec{t}_k\|^2 + \overrightarrow{AC}^2}{2 \overrightarrow{AC}^2}. \end{aligned} \quad (44)$$

Once we determine  $m$ , the original problem can be written as:

$$\overrightarrow{DB^*} = \underset{\vec{v}_{\perp}}{\text{argmin}} \|\overrightarrow{DB} - \overrightarrow{DB^*}\|^2. \quad (45)$$

$\overrightarrow{DB}$  can be orthogonally decomposed according to  $\overrightarrow{AC}$  as:

$$\overrightarrow{DB} = \overrightarrow{DB}_{\parallel} + \overrightarrow{DB}_{\perp}, \quad (46)$$

where  $\overrightarrow{DB}_{\parallel} \parallel \overrightarrow{AC} \parallel \overrightarrow{DB^*}$  and  $\overrightarrow{DB}_{\perp} \perp \overrightarrow{AC}$ .

Thus,  $\|\overrightarrow{DB} - \overrightarrow{DB^*}\|^2 = \|\overrightarrow{DB}_{\perp} - \overrightarrow{DB^*}\|^2 + \overrightarrow{DB}_{\parallel}^2$ , where  $\overrightarrow{DB}_{\parallel}$  is a constant irrelevant to  $B^*$ . Intuitively, the optimal  $\overrightarrow{DB^*}$  must be parallel to  $\overrightarrow{DB}_{\perp}$  since they are both perpendicular to  $\overrightarrow{AC}$  and sharing the same start point  $D$ . Consequently, we have  $\overrightarrow{DB^*} = n \frac{\overrightarrow{DB}_{\perp}}{\|\overrightarrow{DB}_{\perp}\|}$ , where  $n$  denotes

the norm of  $\overrightarrow{DB}^*$ . Following the norm constraints in Eq. 42, we can determine  $n$  as:

$$n = \sqrt{\|\vec{t}_{pa(k)}\|^2 - m^2 \overrightarrow{AC}^2}. \quad (47)$$

Finally, we can calculate  $B^*$  as:

$$\begin{aligned} B^* &= A + \overrightarrow{AB}, \\ &= A + m\overrightarrow{AC} + n \frac{\overrightarrow{DB}_\perp}{\|\overrightarrow{DB}_\perp\|}, \end{aligned} \quad (48)$$

where

$$\overrightarrow{DB}_\perp = \overrightarrow{DB} - \frac{\overrightarrow{DB} \cdot \overrightarrow{AC}}{\|\overrightarrow{AC}\|^2} \overrightarrow{AC}. \quad (49)$$

## APPENDIX C DATASETS

**AGORA [21]:** It is a synthetic dataset featuring precise SMPL and SMPL-X annotations fitted to 3D scans. We use this dataset for training only when conducting experiments on it. The evaluation is performed on the official platform with separated SMPL and SMPL-X tracks.

**3DPW [16]:** It is a challenging outdoor benchmark for body-only 3D pose and shape estimation. It contains 60 video sequences obtained from a hand-held moving camera. It employs IMU sensors to calculate ground-truth pose and shape data.

**MPI-INF-3DHP [18]:** It is a body-only 3D keypoint dataset with both constrained indoor and complex outdoor scenes. It includes 8 actors performing 8 activities from 14 camera views. Following [6], [7], we use its train set for training and evaluate on its test set.

**Human3.6M [17]:** It is an indoor benchmark for body-only 3D pose and shape estimation. It includes 11 subjects performing 15 different activities in a laboratory environment. Following [6], [7], we use 5 subjects (S1, S5, S6, S7, S8) for training and 2 subjects (S9, S11) for evaluation.

**FreiHAND [19]:** It is a single-hand 3D pose dataset with MANO [124] annotations. It contains over 130k training samples with the right hands. We employ HybrIK on the hand pose model and use this dataset for evaluation.

**HO3D-v2 [20]:** It is a dataset with 3D pose annotations for hands and objects under severe occlusions from each other. It contains sequences of the right hand interacting with an object. We use this dataset to evaluate the performance of hand pose estimation under challenging scenarios.

**MSCOCO [117]:** It is a large-scale in-the-wild 2D human pose dataset consisting of over 150k instances. We incorporate its train set for training.

TABLE 12  
**Naive vs. Adaptive** with different input joints on body-only mesh recovery. MPJPE of 24 joints is reported. Adaptive HybrIK is more robust to the noise.

	GT Joints	$\pm 10$ mm	$\pm 20$ mm	$\pm 30$ mm
Naive HybrIK	0.1	43.8	74.1	100.2
Adaptive HybrIK	0.1	21.1	43.6	67.5

## APPENDIX D ABLATION EXPERIMENTS.

**Robustness of HybrIK to Noisy Joint Positions.** In the main paper, we evaluate the robustness of HybrIK in whole-body scenarios. Here, we further report the evaluation of robustness on body-only scenarios. We use the same evaluation protocol as in the main paper. We randomly add Gaussian noise to the 3D joint positions with different standard deviations. The results are shown in Tab. 12. We can see that adaptive HybrIK is more robust to the noise than naive HybrIK.

TABLE 13  
**Reconstruction error with different shape parameters  $\beta$**  on the AGORA validation set.

	GT $\beta$		Estimated $\beta$		Zero $\beta$	
	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE
Error	65.4	74.9	67.9	77.1	73.4	83.4

**Effect of  $\beta$ .** In this experiment, we examine the impact of shape parameters  $\beta$  on the AGORA validation set. As shown in Tab. 13, using the ground-truth  $\beta$  yields a 2 mm improvement in MPJPE and PVE, while using zero  $\beta$  results in a 6 mm error. It shows that our model also gives an accurate body shape estimation.

**Error correction capability of HybrIK.** In this experiment, we investigate the error correction capability of body-only HybrIK on the 3DPW [16], Human3.6M [17], and AGORA [21] datasets. Quantitative results are reported in Tab. 14. They demonstrate that HybrIK can leverage the structural information embedded in the statistical body model to rectify unrealistic body joints derived from off-the-shelf 3D keypoint estimation approaches. The error correction capability of HybrIK is more pronounced on the AGORA dataset, which poses significant challenges for pose and joint estimation.

## APPENDIX E QUALITATIVE RESULTS

Additional qualitative results for body-only, hand-only, and whole-body scenarios are presented in Fig. 9, 10, and 11, respectively. Qualitative comparisons between heatmap-based backbone and our regression-based backbone are displayed in Fig. 12. Qualitative comparisons with state-of-the-art approaches are presented in Fig. 13. More results can be found in our project page <https://jeffli.site/HybrIK-X/>.

TABLE 14  
Error correction capability of HybrIK on the Human3.6M, 3DPW, and AGORA datasets.

	Human3.6M		3DPW		AGORA	
	Predicted Pose	HybrIK	Predicted Pose	HybrIK	Predicted Pose	HybrIK
MPJPE (24 joints) ↓	50.9	48.1	78.9	74.8	73.1	66.3



Fig. 9. Qualitative results of body-only mesh recovery on challenging poses.

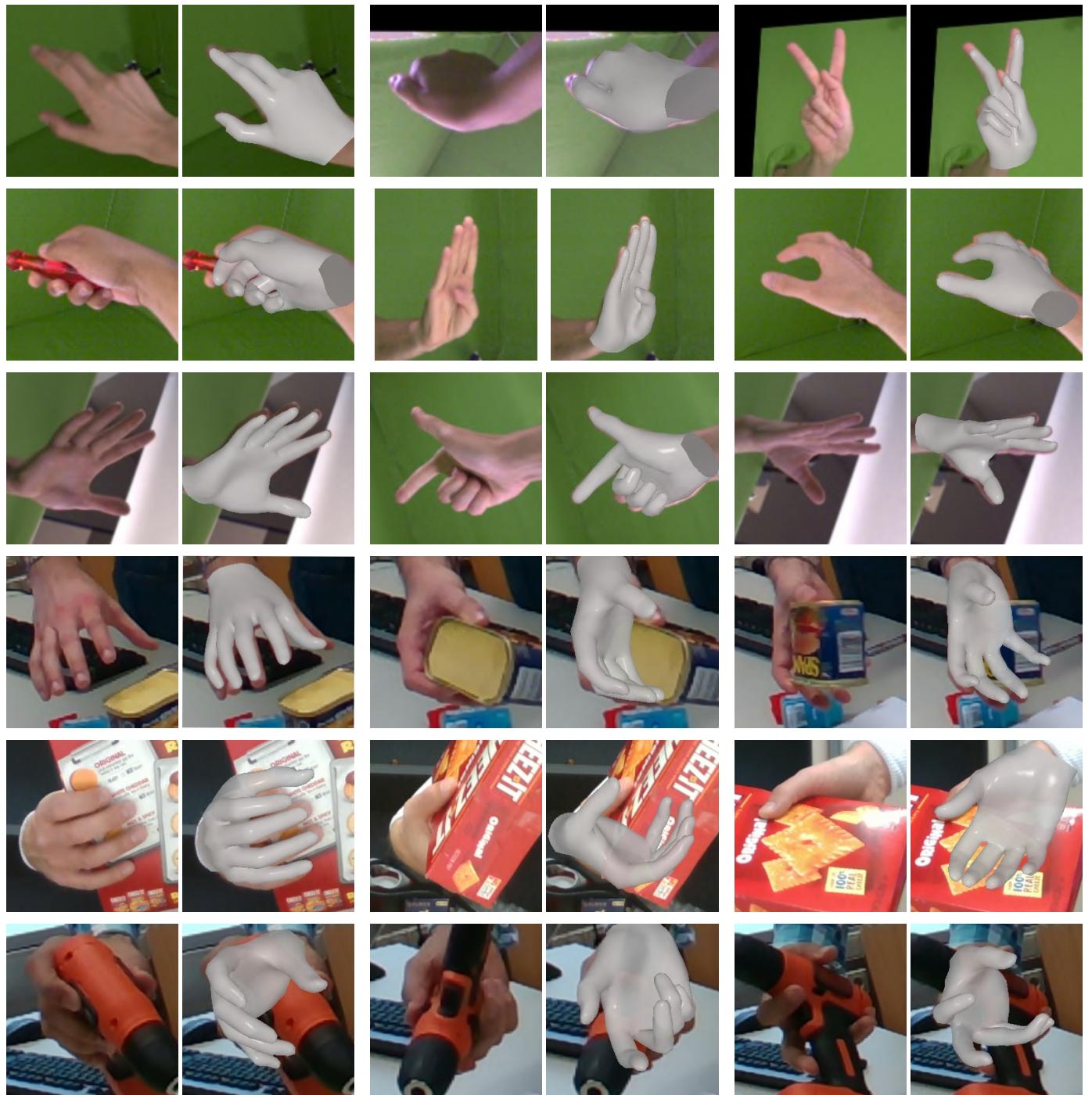


Fig. 10. Qualitative results of hand-only mesh recovery on the FreiHAND dataset (rows 1-3) and the HO3D-v2 dataset (rows 4-6).

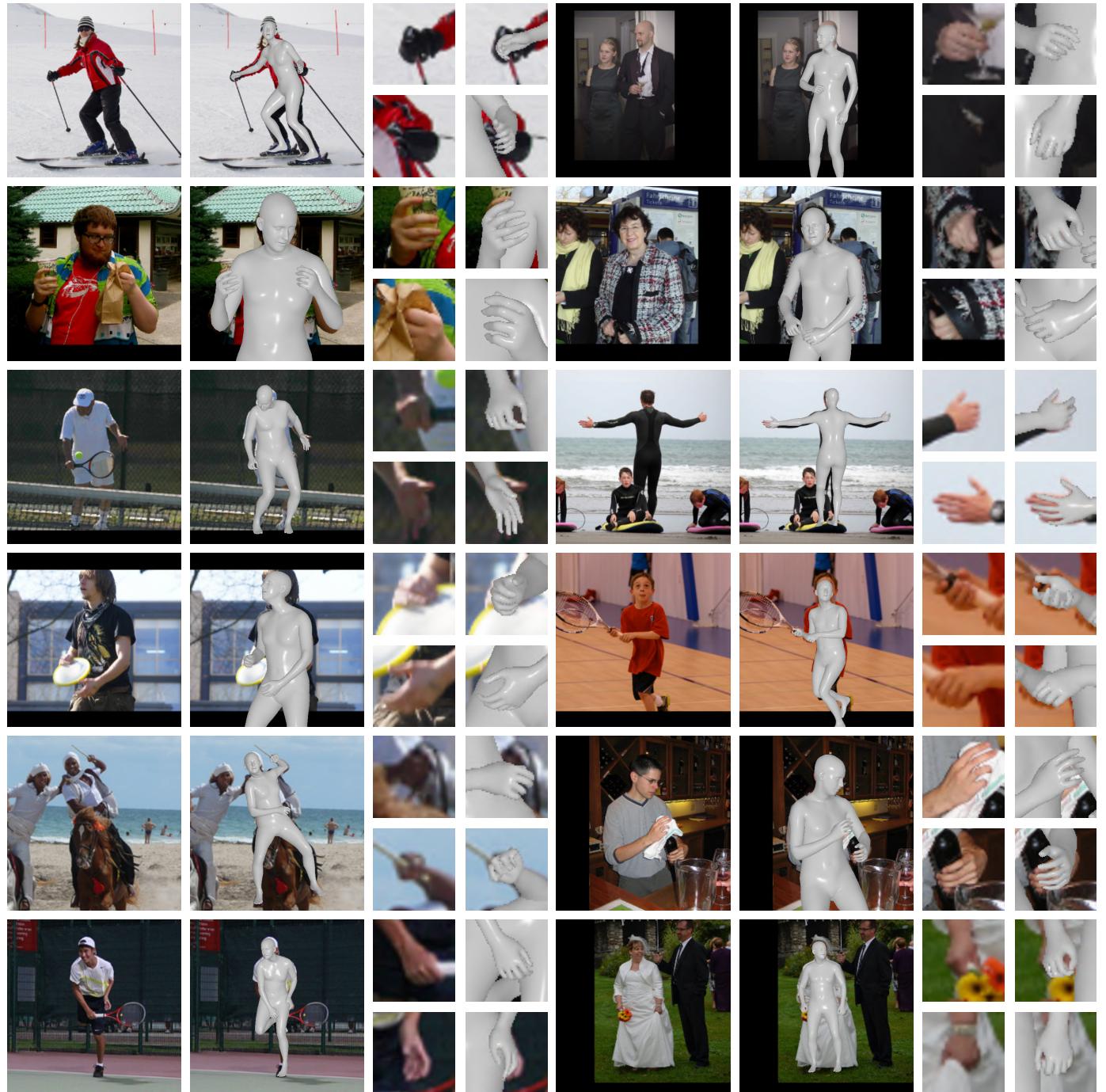


Fig. 11. Qualitative results of whole-body mesh recovery on challenging poses.

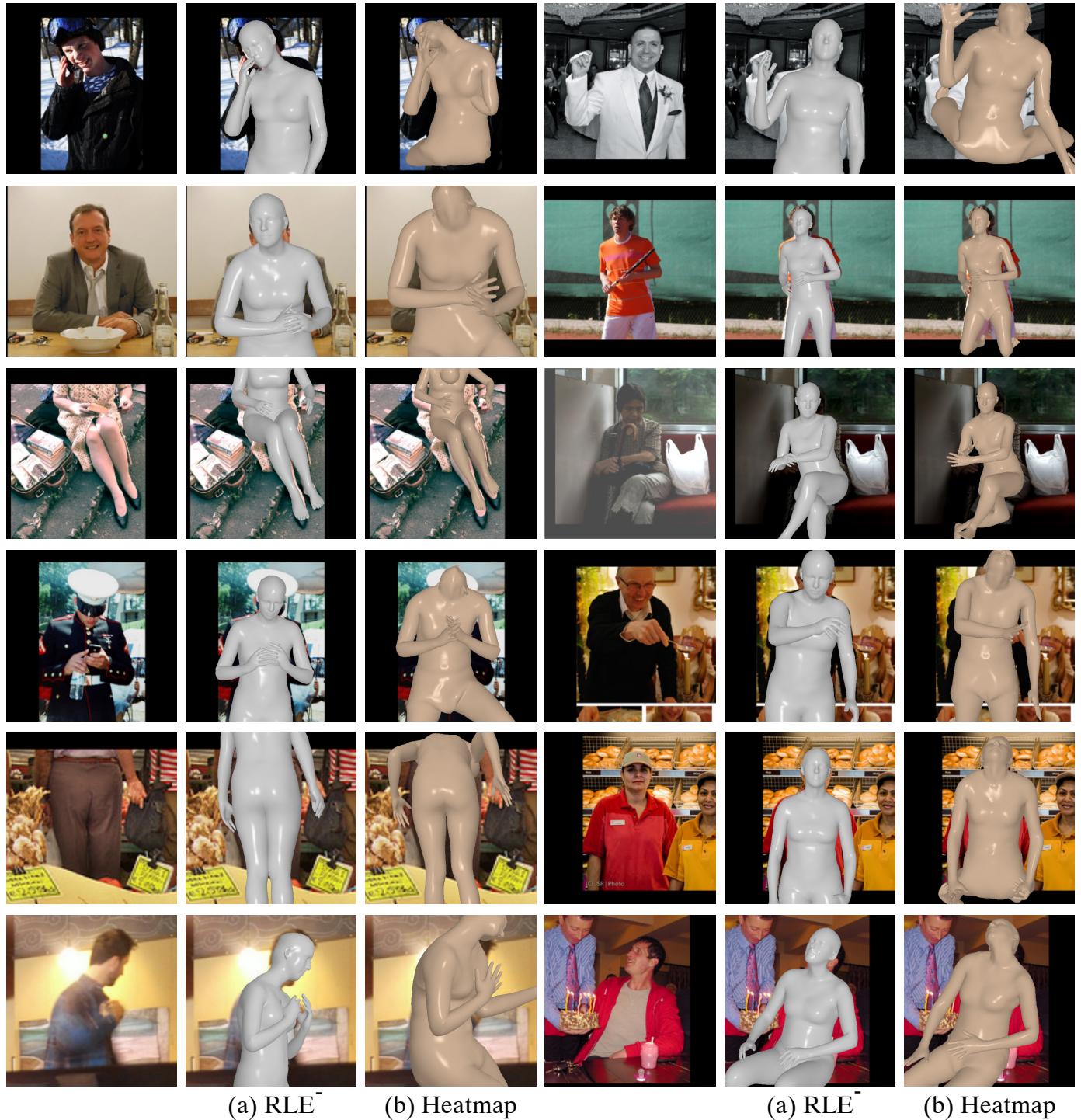


Fig. 12. Qualitative comparisons between RLE<sup>-</sup>-based and heatmap-based body-only mesh recovery.

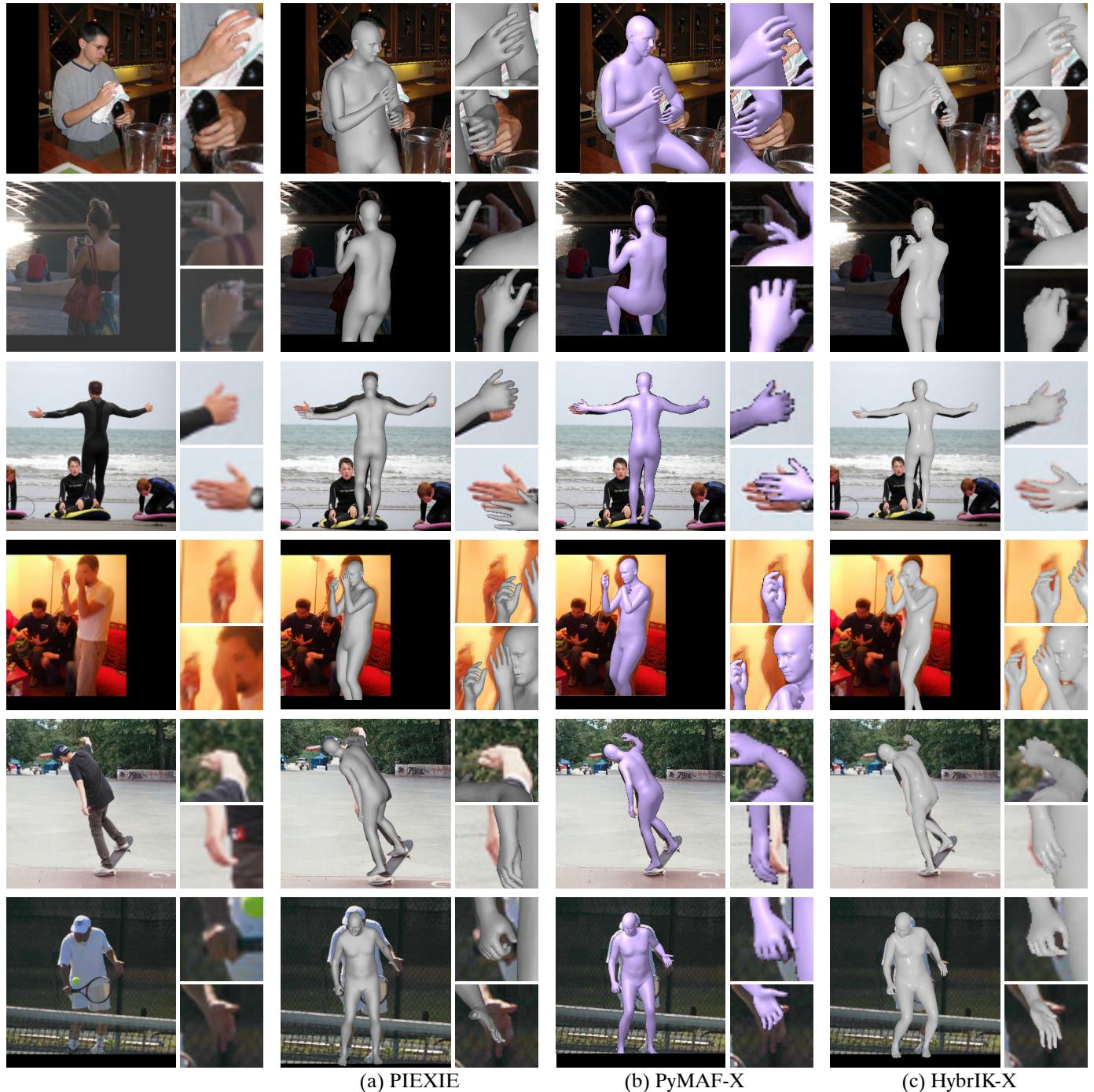


Fig. 13. Qualitative comparisons with state-of-the-art approaches.

## REFERENCES

- [1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *SIGGRAPH*, 2005. 1, 2
- [2] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *TOG*, 2015. 1, 2, 4, 12
- [3] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *CVPR*, 2019. 1, 2, 3, 4, 10, 12
- [4] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *ICCV*, 2009. 1, 2
- [5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*, 2016. 1, 2, 9, 11, 12
- [6] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018. 1, 2, 3, 4, 5, 8, 9, 11, 14
- [7] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019. 1, 2, 3, 4, 5, 8, 9, 11, 14
- [8] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video inference for human body pose and shape estimation," in *CVPR*, 2020. 1, 2, 3, 4, 5
- [9] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *ICCV*, 2021. 1, 2, 3, 9
- [10] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *3DV*, 2021. 1, 2, 8, 10, 12
- [11] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *CVPR*, 2017. 1
- [12] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *ECCV*, 2018. 1
- [13] C. Wang, J. Li, W. Liu, C. Qian, and C. Lu, "Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation," in *ECCV*, 2020. 1, 2
- [14] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," in *ICCVW*, 2021. 2, 3, 10
- [15] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, and Y. Liu, "Pymaf-x: Towards well-aligned full-body model regression from monocular images," *arXiv preprint arXiv:2207.06400*, 2022. 2, 3, 10, 12
- [16] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018. 2, 9, 14
- [17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *TPAMI*, 2014. 2, 9, 14
- [18] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017. 2, 9, 14
- [19] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *ICCV*, 2019. 2, 9, 10, 14
- [20] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *CVPR*, 2020. 2, 9, 10, 14
- [21] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *CVPR*, 2021. 2, 9, 14
- [22] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybrIK: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3383–3393. 2
- [23] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *CVPR*, 2017. 2
- [24] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net: Localization-classification-regression for human pose," in *CVPR*, 2017. 2
- [25] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017. 2
- [26] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *ICCV*, 2017. 2
- [27] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3d pose estimation from monocular rgb," in *3DV*, 2018. 2
- [28] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *ECCV*, 2018. 2
- [29] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *ICCV*, 2019. 2
- [30] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3d human pose reconstruction," in *CVPR*, 2015. 2
- [31] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3d human pose from 2d image landmarks," in *ECCV*, 2012. 2
- [32] H.-Y. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision," in *ICCV*, 2017. 2
- [33] M. Sanzari, V. Ntouskos, and F. Pirri, "Bayesian image based 3d pose estimation," in *ECCV*, 2016. 2
- [34] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3d shape estimation: A convex relaxation approach," *TPAMI*, 2016. 2
- [35] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3d human pose estimation from monocular video," in *CVPR*, 2016. 2
- [36] S. Park, J. Hwang, and N. Kwak, "3d human pose estimation using convolutional neural networks with 2d pose information," in *ECCV*, 2016. 2
- [37] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3d pose estimation from a single image," in *CVPR*, 2016. 2
- [38] F. Moreno-Noguer, "3d human pose estimation from a single image via distance matrix regression," in *CVPR*, 2017. 2
- [39] H. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3d pose estimation," *AAAI*, 2017. 2
- [40] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017. 2
- [41] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *ICCV*, 2017. 2
- [42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014. 2
- [43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 2
- [44] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016. 2
- [45] B. Pang, K. Zha, H. Cao, C. Shi, and C. Lu, "Deep rnn framework for visual sequential applications," in *CVPR*, 2019. 2
- [46] B. Pang, K. Zha, H. Cao, J. Tang, M. Yu, and C. Lu, "Complex sequential understanding through the awareness of spatial and temporal concepts," *Nature Machine Intelligence*, 2020. 2
- [47] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3d human body shapes," in *ECCV*, 2018. 2
- [48] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019. 2
- [49] G. Moon and K. M. Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," in *ECCV*, 2020. 2

- [50] L. Sigal, A. Balan, and M. J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," in *NeurIPS*, 2008. 2
- [51] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *CVPR*, 2017. 2
- [52] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints," in *CVPR*, 2018. 2
- [53] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *CVPR*, 2018. 2, 3, 9
- [54] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *3DV*, 2018. 2, 3
- [55] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *ICCV*, 2021. 2, 3, 9
- [56] Y. Sun, Q. Bao, W. Liu, Y. Fu, M. J. Black, and T. Mei, "Monocular, one-stage, regression of multiple 3d people," in *ICCV*, 2021. 2, 9
- [57] U. Iqbal, K. Xie, Y. Guo, J. Kautz, and P. Molchanov, "Kama: 3d keypoint aware body mesh articulation," in *3DV*, 2021. 2, 9
- [58] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *CVPR*, 2021. 2, 9, 10
- [59] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragiadaki, "Self-supervised learning of motion capture," in *NeurIPS*, 2017. 3
- [60] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," in *CVPR*, 2019. 3
- [61] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3d human reconstruction in-the-wild," in *CVPR*, 2019. 3
- [62] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *CVPR*, 2020. 3
- [63] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large pose 3d face reconstruction from a single image via direct volumetric cnn regression," in *ICCV*, 2017. 3
- [64] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018. 3
- [65] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *CVPR*, 2018. 3
- [66] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *CVPR*, 2019. 3
- [67] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ToG*, 2021. 3
- [68] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng, "End-to-end hand mesh recovery from a monocular rgb image," in *ICCV*, 2019. 3
- [69] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering," in *CVPR*, 2019. 3
- [70] A. Boukhayma, R. d. Bem, and P. H. Torr, "3d hand shape and pose from images in the wild," in *CVPR*, 2019. 3, 10
- [71] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *CVPR*, 2019. 3, 10
- [72] D. Kulon, H. Wang, R. A. Güler, M. Bronstein, and S. Zafeiriou, "Single image 3d hand reconstruction with mesh convolutions," in *BMVC*, 2019. 3
- [73] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *CVPR*, 2019. 3
- [74] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "Inter-hand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *ECCV*, 2020. 3
- [75] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *CVPR*, 2020. 3
- [76] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. Wang, "Interacting two-hand 3d pose and shape reconstruction from single color image," in *ICCV*, 2021. 3
- [77] Y. Rong, J. Wang, Z. Liu, and C. C. Loy, "Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements," in *3DV*, 2021. 3
- [78] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu, "Interacting attention graph for single image two-hand reconstruction," in *CVPR*, 2022. 3
- [79] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *CVPR*, 2018. 3
- [80] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Ghum & ghuml: Generative 3d human shape and articulated pose models," in *CVPR*, 2020. 3
- [81] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3
- [82] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [83] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *CVPR*, 2019. 3
- [84] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *ECCV*, 2020. 3, 10
- [85] Y. Zhou, M. Habermann, I. Habibie, A. Tewari, C. Theobalt, and F. Xu, "Monocular real-time full body capture with inter-part correlations," in *CVPR*, 2021. 3, 10
- [86] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *3DV*, 2021. 3
- [87] G. Moon, H. Choi, and K. M. Lee, "Accurate 3d hand pose estimation for whole-body 3d human mesh estimation," in *CVPRW*, 2022. 3, 9, 10
- [88] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *ECCV*, 2016. 3
- [89] D. Pavillo, D. Grangier, and M. Auli, "Quaternet: A quaternion-based recurrent model for human motion," in *BMVC*, 2018. 3
- [90] Y. Yoshiyasu, R. Sagawa, K. Ayusawa, and A. Murai, "Skeleton transformer networks: 3d human pose and skinned mesh from single rgb image," in *ACCV*, 2018. 3
- [91] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera," *arXiv preprint arXiv:1907.00837*, 2019. 3
- [92] A. Balestrino, G. De Maria, and L. Sciavicco, "Robust control of robotic manipulators," *IFAC Proceedings Volumes*, 1984. 3
- [93] W. A. Wolovich and H. Elliott, "A computational technique for inverse kinematics," in *CDC*, 1984. 3
- [94] M. Girard and A. A. Maciejewski, "Computational modeling for the computer animation of legged figures," in *SIGGRAPH*, 1985. 3
- [95] C. A. Klein and C.-H. Huang, "Review of pseudoinverse control for use with kinematically redundant manipulators," *IEEE Transactions on Systems, Man, and Cybernetics*, 1983. 3
- [96] C. W. Wampler, "Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods," *IEEE Transactions on Systems, Man, and Cybernetics*, 1986. 3
- [97] S. R. Buss and J.-S. Kim, "Selectively damped least squares for inverse kinematics," *Journal of Graphics tools*, 2005. 3
- [98] D. G. Luenberger, Y. Ye et al., *Linear and nonlinear programming*, 1984. 3
- [99] A. Aristidou and J. Lasenby, "Fabrik: A fast, iterative solver for the inverse kinematics problem," *Graphical Models*, 2011. 3
- [100] N. Rokbani, A. Casals, and A. M. Alimi, "Ik-fa, a new heuristic inverse kinematics solver using firefly algorithm," in *Computational Intelligence Applications in Modeling and Control*, 2015. 3
- [101] D. Tolani, A. Goswami, and N. I. Badler, "Real-time inverse kinematics techniques for anthropomorphic limbs," *Graphical models*, 2000. 3, 5
- [102] M. Kallmann, "Analytical inverse kinematics with body posture control," *Computer animation and virtual worlds*, 2008. 3, 5
- [103] A. Csiszar, J. Eilers, and A. Verl, "On solving the inverse kinematics problem using neural networks," in *M2VIP*, 2017. 3
- [104] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *CVPR*, 2018. 3

- [105] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *ICCVW*, 2017. [3](#)
- [106] M. Kopic, D. Kratic, and J. Bohg, "Learning to estimate pose and shape of hand-held objects from rgb images," in *IROS*, 2019. [3](#)
- [107] P. Baerlocher and R. Boulic, "Parametrization and range of motion of the ball-and-socket joint," in *Deformable avatars*. Springer, 2001. [3](#)
- [108] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming, "Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary," Sytronics Inc Dayton Oh, Tech. Rep., 2002. [4](#)
- [109] N. Kofinas, E. Orfanoudakis, and M. G. Lagoudakis, "Complete analytical inverse kinematics for nao," in *ICARSC*, 2013. [5](#)
- [110] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 025–11 034. [8](#)
- [111] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019. [8](#)
- [112] G. Moon and K. M. Lee, "I2L-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," in *ECCV*, 2020. [9](#), [10](#)
- [113] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation," in *3DV*, 2021. [9](#)
- [114] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, "Spec: Seeing people in the wild with an estimated camera," in *ICCV*, 2021. [9](#)
- [115] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3d people in depth," in *CVPR*, 2022. [9](#)
- [116] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *ECCV*, 2022. [9](#)
- [117] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014. [9](#), [14](#)
- [118] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *ECCV*, 2020. [10](#)
- [119] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. Tan, "I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling," in *ICCV*, 2021. [10](#)
- [120] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *CVPR*, 2020. [10](#)
- [121] K. Li, L. Yang, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu, "Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis," in *CVPR*, 2021. [10](#)
- [122] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation," in *CVPR*, 2022. [10](#)
- [123] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ToG*, 2017. [10](#)
- [124] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *TOG*, 2017. [14](#)