3D Human Mesh Recovery with Sequentially Global Rotation Estimation

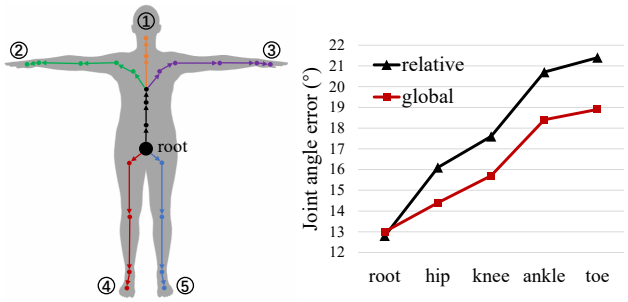
Dongkai Wang Shiliang Zhang
National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{dongkai.wang, slzhang.jdl}@pku.edu.cn

Abstract

Model-based 3D human mesh recovery aims to reconstruct a 3D human body mesh by estimating its parameters from monocular RGB images. Most of recent works adopt the Skinned Multi-Person Linear (SMPL) model to regress relative rotations for each body joint along the kinematics chain. This pipeline needs to transform each relative rotation matrix into a global rotation matrix to articulate the canonical mesh, and suffers from accumulated errors along the kinematics chain. This paper proposes to directly estimate the global rotation of each joint to avoid error accumulation and pursue better accuracy. The proposed Sequentially Global Rotation Estimation (SGRE) directly predicts the global rotation matrix of each joint on the kinematics chain. SGRE features a residual learning module to leverage complementary features and previously predicted rotations of parent joints to guide the estimation of subsequent child joints. Thanks to this global estimation pipeline and residual learning module, SGRE alleviates error accumulation and produces more accurate 3D human mesh. It can be flexibly integrated into existing regression-based methods and achieves superior performance on various benchmarks. For example, it improves the latest method 3DCrowdNet by 3.3 mm MPJPE and 5.0 mm PVE on 3DPW dataset and 3.0 AP on COCO dataset, respectively[†].

1. Introduction

3D human mesh recovery aims to estimate the 3D surface mesh of a human body from monocular RGB images. It has a wide range of applications in human-object interaction, action recognition and virtual/augmented reality. Thanks to parametric human body models [2, 29, 33], 3D human mesh recovery can be simplified, *i.e.*, a realistic 3D mesh of human body can be generated by a few parameters like shape parameters and joint rotations. Most of recent 3D human mesh recovery methods can be regarded as the model-based method, which estimates shape and rotation parameters to recover the 3D human mesh.



(a) (b)

Figure 1. (a) Illustration of the five kinematics chains in SMPL model. (b) Visualization of the joint rotation angle error along one of kinematics chain from root to toe. It can be observed that error increases along the chain, and our global estimation gets more accurate joint rotation than the relative estimation.

Current model-based human mesh recovery methods can be summarized into two categories according to their followed pipelines. Optimization-based approaches [4, 25] estimate the body pose and shape parameters by an iterative fitting process. Parameters of the statistical model are tuned to reduce the error between its 2D projection and 2D evidences, *e.g.*, 2D joint locations and silhouette, which can be obtained by current advanced methods [5, 40, 38, 39]. These methods can typically produce well-aligned results, but could take a long time because the optimization is non-convex. These methods are also sensitive to the initialization. Regression-based methods adopt the powerful neural networks to directly regress model parameters [18, 22, 7], which have exhibited promising results. To tackle the difficulty of non-linear mapping from input image to parameter space, many regression-based methods have been proposed [27, 15, 42]. More detailed review of existing methods can be found in Sec. 2.

Existing regression-based methods regress the relative rotation matrix for each joint with respect to (*w.r.t*) their parent joints along the kinematics chain as illustrated in Fig. 1 (a). This design is partially because parametric body

[†]Code & Model: https://github.com/kennethwdk/SGRE

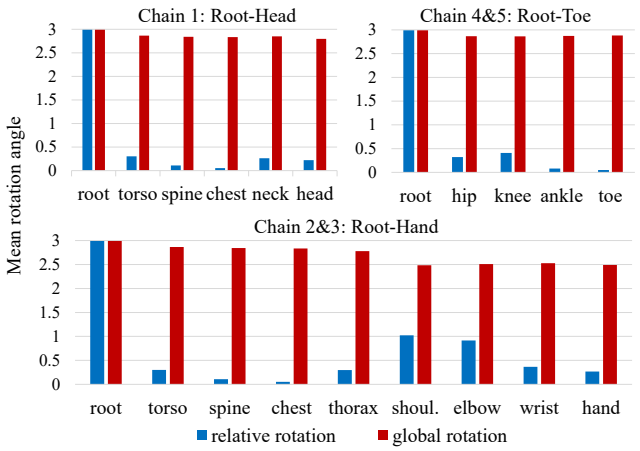


Figure 2. Visualization of the average angle of relative and global rotation of 5 kinematics chains, *e.g.*, root-head, root-toe and root-hand, where parent and child joints are highly correlated, *i.e.*, the relative rotation angle is substantially smaller. The results are obtained on 3DPW dataset.

models like SMPL [29] adopt relative rotations as parameters. To articulate the canonical mesh, those models need to transform each relative rotation matrix into a global rotation matrix by multiplying its parent rotation matrixes along the kinematics chain. This procedure propagates and accumulates errors in parent rotation matrices into those of child joints, leading to larger rotation errors long the kinematics chain. As illustrated in Fig. 1 (b), larger errors can be observed in the end of the kinematics chain, which represents end joints like head and ankle that are far from the root joint.

Instead of predicting relative rotations, this paper proposes to directly estimate the global rotation for each joint through end-to-end optimization *w.r.t.* the groundtruth, to alleviate the accumulated errors. Compared with relative rotation estimation, the global rotation estimation is more challenging because of a larger solution space. In other words, the relative rotation angle of each joint is commonly small *w.r.t.* its parent joints, as illustrated in Fig. 2. Differently, global rotation exhibits larger freedom of rotation angles, making an accurate prediction more difficult.

To relieve the difficulty of direct global estimation, we leverage the rotation of parent joints to estimate the rotation of child joints as they are closely correlated as shown in Fig. 2. This intuition leads to the Sequentially Global Rotation Estimation (SGRE), which leverages rotation matrixes of some joints as references to guide the estimation of subsequent joints on the kinematics chain, a similar way of residual learning [13]. Considering that rotation can be determined by the corresponding body part with canonical pose [15], SGRE further imposes the local body part features. Experiments show that, SGRE effective alleviates error accumulation and produces more accurate 3D reconstruction results for end joints like neck, head, ankle, *etc.*

SGRE presents a novel direct rotation prediction pipeline. It is general and can be flexibly integrated into existing regression-based methods by replacing their relative rotation estimation branches. We test the effectiveness of SGRE on both a baseline regression model and the latest 3DCrowdNet [7] with extensive experiments on various 3D/2D human pose and shape estimation benchmarks. Integrating 3DCrowdNet [7] with SGRE, we achieve 78.4 MPJPE and 93.3 PVE on 3DPW [37] test set, outperforming the original 3DCrowdNet by 3.3 MPJPE and 5.0 PVE. On 3DPW-OCC dataset, SGRE exhibits more substantial advantages and reduces joint errors by 4.0 MPJPE.

To the best of our knowledge, this is an original effort on global rotation estimation for 3D human mesh recovery. Compared with the relative rotation estimation pipeline followed by previous works, SGRE effective alleviates the error accumulation issue and produces better results. The SGRE can be flexibly applied to recent regression-based 3D human mesh recovery works to further boosts their performance. Those advantages make SGRE a promising method for 3D human mesh recovery.

2. Related Work

This part briefly reviews existing 3D human mesh recovery works, that can be summarized into model-based and model-free methods, respectively.

Model-based 3D human mesh recovery adopts parametric human model like SMPL [29] and estimates its parameters to generate 3D human mesh. Because parameters of human model are embedded in the low dimensional space and provide a strong structure prior of human body, model-based methods can be trained with few annotations, *e.g.*, 2D evidences. Benefited by this property, model-based methods have dominated this area. Related works can be divided into optimiztion-based and regression-based ones.

Optimization-based methods: Due to the lack of large 3D annotation datasets, early model-based 3D mesh recovery methods follow an iterative optimization process. Parameters of the statistical model are tuned to reduce the error between 2D projection and 2D evidences, such as 2D joint locations and silhouette [11, 4]. The objective function typically contains a regularization part to penalize the unnatural shape and pose, as well as another part to measure the fitting error between the 2D projection and 2D evidences. SMPLify [4] detects 2D keypoints with off-the-shelf CNNs as evidence to iteratively train model parameters. 3D body joints [44], silhouettes [25] and part segmentation [41] are also adopted in some works as evidence. HoloPose [12] refines the regressed model parameters with FCN estimated DensePose, 2D and 3D keypoints. Despite their well-aligned results, those method need a slow training procedure and are sensitive to initialization.

Regression-based methods [18, 23, 16, 6, 22, 7] take ad-

vantage of the powerful nonlinear mapping capability of neural networks, *e.g.*, CNN [13] or Transformer [9] to directly predict pose and shape parameters of parametric human model from 2D images. These methods learn model priors implicitly in a data-driven manner under different types of supervision including parameters loss, 3D joints loss and 2D projection loss, *etc.* To relieve the difficulty of directly regression, researchers have proposed many methods. A line of works progressively refines the regressed results in a loop. HMR [18] adopts an adversarial prior and an Iterative Error Feedback (IEF) loop to reduce the difficulty of regression. PyMAF [42] further addresses the misalignment between estimated mesh and input evidence in IEF and proposes a mesh alignment feedback. Another line of works proposes powerful networks to regress model parameters. PARE [22] designs a part attention module to enhance the model capability and handle occlusion. 3DCrowdNet [7] proposes a joint-based regressor with graph convolution to better estimate model parameters.

Recently there appears the third line of works that adopts inverse kinematics to analytically estimate pose parameters from 3D keypoints. KAMA [15] uses a learned 3D keypoint model to obtain the 3D location of each joint. Then, the rotation of each part can be analytically calculated. An optional refinement can be applied for shape estimation and pose refinement. HybrIK [27] decouples the rotation into twist and swing parts, then estimates twist rotation with a network and analytically calculates the swing rotation based on estimated 3D keypoints. Those methods relieve the difficulty of rotation regression and achieve superior performance. There also exist other methods to improve the performance of regression-based methods. SPIN [23] adds an optimization step after regression to provide extra 3D supervision from unlabeled images. Temporal context information [3, 21] is also exploited for better regression.

Model-free 3D human mesh recovery directly estimates the mesh vertex coordinates instead of parameters of parametric human model. Those methods commonly rely on pseudo-groundtruth mesh vertex annotations on large-scale datasets [14, 28, 37]. Among those methods, GraphCMR [24] adopts graph convolutional neural network to directly predict mesh vertex location. Pose2Mesh [6] adopts graph network to recover 3D human pose and mesh from 2D pose. I2L-MeshNet [32] proposes a lixel 1D heatmap to encode the 3D location of each mesh vertex on x, y, z dimension separately.

Difference with previous works: This paper proposes a novel direct rotation prediction pipeline. It alleviates the accumulated error in the relative rotation prediction pipeline commonly used in existing methods. The proposed method can be integrated with existing regression-based methods to further boost the performance. Therefore, this work differs with existing works in both motivation and methodology.

3. Method

3.1. Overview

The goal of 3D human mesh recovery is to estimate the triangulated mesh $\mathbf{M} \in \mathbb{R}^{N\times 3}$ with $N = 6980$ vertices from monocular input RGB image $\mathcal{I}$, which can be conceptually denoted as

$$\mathbf{M} = \text{Recovery}(\mathcal{I}). \tag{1}$$

We leverage the SMPL model to compute $\mathbf{M}$. SMPL [29] is a parametric human body model that allows to use shape and pose parameters to reconstruct the 3D human body mesh. The shape parameters $\beta \in \mathbb{R}^{10}$ are the first 10 principal components of the shape space. The pose parameters θ are 3D rotation matrixes of K=24 joints. We denote it as $\theta = (\theta_1, \theta_2, ..., \theta_K)$, where θ_1 is the global rotation of root joint, and θ_k $(k > 1)$ denotes the relative rotation of the k-th joint *w.r.t* its parent. The parent-child relation is defined by the kinematics chain illustrated in Fig. 1 (a). SMPL provides a differentiable function $\mathbf{M} = \mathcal{M}(\theta, \beta)$ that outputs the mesh $\mathbf{M}$. With $\mathbf{M}$, the coordinates of 3D joint $\mathcal{J}_{3D} \in \mathbb{R}^{K\times 3}$ can be obtained by applying a pre-trained linear regressor on $\mathbf{M}$. More details about SMPL can be found in [29] and previous regression-based methods [18, 7].

With SMPL, our goal is to estimate pose and shape parameters by training a CNN model $\Phi(\cdot)$, *i.e.*, $\{\theta, \beta\} = \Phi(\mathcal{I})$, where $\{\theta, \beta\}$ are used to generate the 3D body mesh with the function $\mathcal{M}(\theta, \beta)$. More specifically, the backbone $\Phi(\cdot)$ takes an image $\mathcal{I}$ as input and outputs a feature map $\mathcal{F} \in \mathbb{R}^{C\times H\times W}$, where H and W denote the spatial size of feature map. Following HMR [18], we apply global average pooling on $\mathcal{F}$ to get a global feature vector f_g. Then a parameter regressor $\mathcal{R}(\cdot)$ takes feature f_g as input and outputs the estimated parameters θ and β, respectively, *i.e.*,

$$\theta = \mathcal{R}_{\texttt{pose}}(f_g), \beta = \mathcal{R}_{\texttt{shape}}(f_g), \tag{2}$$

where parameters θ and β are feed into SMPL to generate mesh vertices $\mathbf{M} = \mathcal{M}(\theta, \beta)$. We illustrate this inference procedure in Fig. 3.

To train the backbone $\Phi(\cdot)$, we map mesh vertices to 3D joints $\mathcal{J}_{3D} \in \mathbb{R}^{K\times 3}$ by the pretrained linear regressor in SMPL, then further project 3D joints to the image coordinate system as 2D keypoints $\mathcal{J}_{2D} = \mathbf{\Pi}(\pi, \mathcal{J}_{3D})$, where $\mathcal{J}_{2D} \in \mathbb{R}^{K\times 2}$ and $\mathbf{\Pi}(\phi, \cdot)$ denotes the projection function based on the camera parameters π predicted via a regressor $\pi = \mathcal{R}_{\texttt{cam}}(f_g)$.

As minor parameter errors in 3D space can lead to large misalignment in 2D space, we follow previous works [18, 22, 7] to add supervision on the projected 2D keypoints to penalize 2D misalignment. Meanwhile, additional 3D supervision on 3D joints and SMPL parameters are also considered during the training. The overall training loss can be

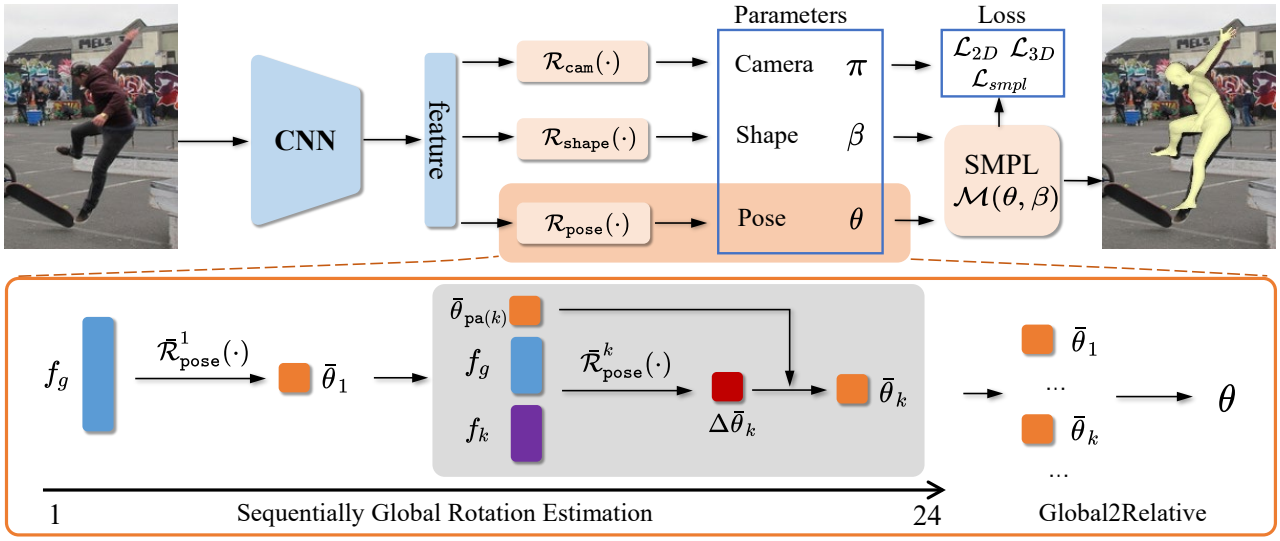


Figure 3. The proposed framework for 3D human mesh recovery. A backbone, *e.g.*, CNN is adopted to extract the image feature, which is sent to different regressors to obtain the camera, shape and pose parameters to generate final mesh with the SMPL model. The complicated pose parameter θ is estimated by Sequentially Global Rotation Estimation (SGRE) to alleviate accumulated errors in previous methods.

denoted as,

$$\mathcal{L} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{SMPL}\mathcal{L}_{SMPL}, \quad (3)$$

where the superscript "*" denotes the groundtruth, and each term is calculated as

$$\mathcal{L}_{2D} = ||\mathcal{J}_{2D} - \mathcal{J}^*_{2D}||^1, \quad (4)$$

$$\mathcal{L}_{3D} = ||\mathcal{J}_{3D} - \mathcal{J}^*_{3D}||^1, \quad (5)$$

$$\mathcal{L}_{SMPL} = ||\theta - \theta^*||^2 + ||\beta - \beta^*||^2. \quad (6)$$

Fig. 3 illustrates our framework, which regresses three parameters ϕ, β, θ. Among those parameters, the pose parameter $\theta = (\theta_1, \theta_2, ..., \theta_K)$ is the most complicated one and largely determines the reconstruction accuracy. The following part presents details of our Sequentially Global Rotation Estimation (SGRE), which implements the regressor $\mathcal{R}_{\text{pose}}(\cdot)$.

3.2. Sequentially Global Rotation Estimation

Relative Rotation Estimation: Previous methods estimate the relative rotations from input image via deep neural networks, which can be denoted as,

$$\{\theta_k\}_{k=1}^K = \mathcal{R}_{\text{pose}}(f_g), \quad (7)$$

where θ_k denotes the relative rotation of k-th joint. To articulate the mesh, the global rotation $\bar{\theta}_k$ of the k-th joint is computed as,

$$\bar{\theta}_k = \prod_{i=1:k} \theta_i. \quad (8)$$

The relative rotation estimation is straightforward and θ can be directly used in SMPL model. However, it suffers from error accumulation along the kinematics chain as illustrated in Eq. (8).

Global Rotation Estimation: We propose to directly estimate the global rotations from the input image, and train the regressor $\bar{\mathcal{R}}_{\text{pose}}(\cdot)$, *i.e.*,

$$\bar{\theta} = \{\bar{\theta}_k\}_{k=1}^K = \bar{\mathcal{R}}_{\text{pose}}(f_g), \quad (9)$$

$$\theta_k = \bar{\theta}^{-1}_{\text{pa}(k)}\bar{\theta}_k, \quad (10)$$

where $\text{pa}(k)$ is the index of parent joint of the k-th joint. $\bar{\theta}^{-1}_{\text{pa}(k)}$ denotes the inverse matrix of parent joint, which can be efficiently computed as the transposed matrix $\bar{\theta}^t_{\text{pa}(k)}$ due to the orthogonality of rotation matrix. We adopt the commonly used 6D representation [45] to represent a rotation, therefore the dimension of $\bar{\theta}$ is $24 \times 6 = 144$.

Because existing datasets only provide the relative rotation annotation, we transform the ground truth relative rotation θ^* into global rotation $\bar{\theta}^*$ via Eq. (8). During inference, we use Eq. (10) to transform the estimated global rotation into relative rotation, which is required by the SMPL model to generate mesh. Our method is optimized *w.r.t.* the global rotation and directly predicts the global rotation. Compared with accumulating multiple relative predictions to get the global rotation as in Eq. (8), our method has better potentials to alleviate the accumulated error during inference.

As shown in Fig. 2, global rotation exhibits larger freedom of rotation, making global prediction more difficult. To address this issue, we introduce two components into global regressor $\bar{\mathcal{R}}_{\text{pose}}(\cdot)$. The first is a sequential estimation procedure that leverages the prediction of parent joints to guide the estimation of child joints. This procedures leverage the

correlation between parent and child joints to relieve the difficulty of directly estimation. The second is a local feature enhancement module that provides extra local part cues as a complementary feature to the commonly used global feature. Those two components constitute the SGRE model.

Sequential Estimation. As shown in Fig. 3, the global rotation of root joint will be computed first. To compute $\bar{\theta}_k$, SGRE takes the rotation matrix of its parent joint, *i.e.*, $\bar{\theta}_{\mathtt{pa}(k)}$ as reference. Therefore, SGRE sequentially estimates the global rotation of each joint along the kinematics chain. As adjacent parts are highly correlated, the global rotation of k-th joint should be constraint by the rotation of its parent joint. We hence predict $\bar{\theta}_k$ as,

$$\bar{\theta}_k = \bar{\mathcal{R}}^k_{\mathtt{pose}}(F_k, \bar{\theta}_{\mathtt{pa}(k)}) + \bar{\theta}_{\mathtt{pa}(k)}, \tag{11}$$

where F_k denotes the visual feature adopted by SGRE to predict $\bar{\theta}_k$.

Eq. (11) differs with the relative rotation estimation of Eq.(7) in several aspects. First, it adopts both the $\bar{\theta}_{\mathtt{pa}(k)}$ and the visual feature as input, hence allows to leverage the rotation of parent joint to guide the rotation prediction of child joint. Secondly, the regressor $\bar{\mathcal{R}}(\cdot)$ in Eq. (11) predicts the residual matrix between k-th joint and its parent, hence is easier than the direct estimation. In other words, this residual learning strategy can output reasonable rotation matrix even the regressor $\bar{\mathcal{R}}(\cdot)$ produces zero output, because $\bar{\theta}_{\mathtt{pa}(k)}$ and $\bar{\theta}_k$ are correlated as shown in Fig. 2.

Local Feature Enhancement. Besides introducing new regressor model, SGRE also adopts complementary features to pursue a better performance. Relative rotation computation involves two adjacent body parts, and requires a large receptive field, especially for large body parts like thigh and calf. Differently, global rotation can be determined by a local body part with canonical pose [15], therefore a local body part clue is more preferred for global rotation estimation. We propose the local feature enhancement to provide local feature of the k-th joint into its global rotation prediction. The visual feature F_k adopted by the SGRE can be denoted as,

$$F_k = \{f_g, f_k\}, \tag{12}$$

where f_k denotes the feature sampled around the location of k-th joint via bilinear interpolation. It is concatenated with the global feature f_g as the F_k to provide complementary global and local visual information.

We adopt the commonly used soft-argmax [34] to obtain the spatial coordinates of each joint from input image. Specifically, given the feature map $\mathcal{F}$ outputted by the backbone, we pass it to several convolutional layers to obtain the estimated heatmap $\mathcal{H}_k$ for k-th joint, then an integration operation is applied to get the continuous coordinates p_k, which is used to sample f_k from $\mathcal{F}$. During training we add L_1 loss to p_k to supervise the keypoint estimation training.

Discussions: SGRE alleviates the difficulty of directly global rotation estimation by estimating the residual rotation of parent and child joint, which is different from relative rotation in optimization objective. SGRE can tolerate the error of parent joint because it only serves as a good starting point in optimization. Errors in starting points could be corrected by the training procedure. While errors in multiple relative rotations of previous works can not be corrected by Eq. (8), and leads to error accumulation during inference.

The residual rotation estimation in SGRE is also different from previous HMR [18] and HKMR [10]. HMR and HKMR update each joint recursively to pursue more accurate relative rotation, where each iteration aims to decrease the residual between the previous prediction and the groundturth of each joint. SGRE explores the kinematics chain to reduce the difficulty of directly estimating global rotation. It estimates the residual rotations between each joint and its parent node.

4. Experiments

4.1. Datasets and Evaluation Metric

Training sets. Following the setting of previous work [7], the proposed method is trained on a mixture of data from several datasets with 3D and 2D annotation, including Human3.6M [14], MuCo-3DHP [31], COCO [28] and MPII [1]. Human3.6M [14] and MuCo-3DHP [31] are large-scale 3D human pose estimation benchmarks. COCO [28] and MPII [1] contain large-scale in-the-wild images with 2D human joint coordinates annotations. We use the pseudo SMPL groundtruth annotation from [7].

Evaluation sets. 3DPW [37] contains 60 video sequences captured mostly in outdoor conditions. We use this dataset only for evaluation on its test set. Following PARE [22], we also report the performance on 3DPW-OCC, an occlusion subset of 3DPW dataset to evaluate the performance of methods under heavy occlusion. To evaluation the alignment of predicted mesh with 2D evidences, we also follow [42, 19] to report the 2D human pose estimation performance on COCO [28] val set and CrowdPose [26] test set by projecting the estimated 3D joints on the image plane.

Evaluation metric. We report 3D pose and shape evaluation metrics, as well as 2D pose estimation metrics. For the 3D evaluation, we use mean per-joint position error (MPJPE), Procrustes-aligned mean per-joint position error (PA-MPJPE) and mean per-vertex position error (MPVPE). All errors are measured after aligning root joints of GT and estimated human body meshes. We also adopt mean per-joint angle error (MPJAE) to evaluate the accuracy of estimated joint rotation. For 2D pose estimation, we report average precision (AP) with different thresholds, object sizes and CrowdIndex [26].

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Relative baseline	88.4	56.7	105.5
Global baseline	87.8	56.3	104.4
SGRE	85.5	55.0	101.8
Relative++	87.6	56.1	105.9

Table 1. Ablation study with other baselines on 3DPW test set.

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Global baseline	87.8	56.3	104.4
+Sequential	86.5	55.6	102.6
+Local feat	85.5	55.0	101.8

Table 2. Validity of each component in SGRE on 3DPW test set.

Infer (s)	Feat Extract	Param Regress	Mesh Generate
Baseline	0.102	0.002	0.045
SGRE	0.102	0.010	0.045

Table 3. Runtime analysis of each component in baseline and SGRE. The inference time is tested with batch size 64 on 3DPW test set.

Method	3DPW MPJAE ↓	3DPW-OCC MPJAE ↓
Relative baseline	22.3	25.3
SGRE	21.0	23.8

Table 4. Analysis on joint rotation error in $SO(3)$, the units of rotation error are in *degree*.

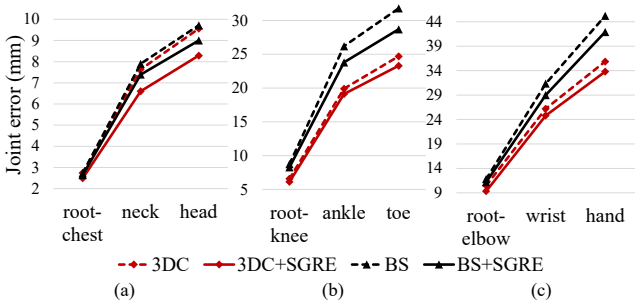


Figure 4. Visualization of the joint error along difference kinematics chains of different rotation estimation methods. (a) Root to head, (b) Root to toe, and (c) Root to hand. BS and 3DC denote the baseline and 3DCrowdNet [7] respectively.

4.2. Implementation Details

All experiments are conducted using PyTorch. We adopt ResNet-50 [13] as backbone for all experiments and follow all the training and test configuration of [7]. The input image is resized to 256× 256 and produces image feature map with size $2048 \times 8 \times 8$. Average pooling is applied on the output feature map and results in a 2048×1 feature vector f_g. We adopt Adam optimizer [20] with a batch size of 160. The learning rate is set to 10^{-4} for backbone and 10^{-3} for other parameters. For ablation study we train model on COCO for 10 epochs and evaluate on 3DPW test set. To compare with other methods, we incorporate all training data and train model for 10 epochs.

4.3. Ablation Study

This section aims to evaluate the effectiveness of SGRE and investigate the contribution of each proposed components. All evaluation is conducted on the 3DPW test set.

Comparison with baselines. We first compare the proposed global rotation estimation with relative rotation estimation under the same setting, *e.g.*, same input and backbone. The results are shown in Table 1. We can observe that directly estimate global rotation achieves similar performance to relative baseline. This demonstrates that estimating global rotation is feasible and can achieve good performance. However, directly estimating global rotation cannot provide accurate results due to the large joint angle. The proposed SGRE can relieve this problem and constantly improve the performance of global rotation estimation. We also evaluate a relative rotation estimation model with sequentially estimation and local feature enhancement, which is denoted as Relative++. It achieves slightly better performance to baseline but is inferior to SGRE. This demonstrates that the proposed SGRE is better for global rotation estimation and alleviate the error accumulation.

Component Analysis. We then analyze the effectiveness of each components. The results are shown in Table 2. We first test the effectiveness of sequentially estimation. Compared with directly estimation, it improves the MPJPE by 1.3 and PVE by 1.8. Local feature enhancement is also useful, which further improves the MPJPE from 86.5 to 85.5 mm. These results demonstrate the effectiveness of each component in SGRE.

Error Analysis. We directly measure the joint rotation error in $SO(3)$ using geodesic distance and the results are shown in Table 4. We can observe that SGRE obtains more accurate joint rotation than baseline. We also conduct error analysis on the joint position along the kinematics tree to investigate the error accumulation phenomenon. The results are shown in Fig. 4. We test three paths in kinematics tree, *e.g.*, from root joint to head, toe and hand joints. The chains are truncated to better illustrate the error of end joints. Fig. 4 shows that the error increases along all kinematics chains. And we can observe that global rotation estimation reduces more error in the deeper joint than relative rotation estimation. This indicates that SGRE can alleviate the error accumulation in relative rotation estimation methods.

Runtime Analysis. Table 3 gives a detailed analysis on the efficiency of each component of SGRE with baseline. The difference is that SGRE sequentially decodes parameters, which is slower than parallel estimation in baseline, *e.g.*, from 0.002 to 0.010. Compared with feature extraction

Method	3DPW		
	MPJPE ↓	PA-MPJPE ↓	PVE ↓
HMMR [18]	116.5	72.6	-
Doersch *et al*. [8]	-	74.7	-
Sun *et al*. [36]	-	69.5	-
VIBE [21]	93.5	56.5	113.4
MEVA [30]	86.9	54.7	-
HMR [18]	130.0	76.7	
CMR [24]	-	70.2	-
SPIN [23]	96.9	59.2	135.1
HMR-EFT [17]	-	54.2	-
Pose2Mesh [6]	89.2	58.9	-
I2L-MeshNet [32]	93.2	58.6	-
ROMP [35]	91.3	54.9	108.3
PyMAF [42]	92.8	58.9	110.1
PARE [22]	82.9	52.3	99.7
3DCrowdNet [7]	81.7	51.5	98.3
Baseline	83.1	52.3	99.9
+SGRE	81.5	51.4	97.4
3DCrowdNet†	82.0	51.4	98.0
+SGRE	**78.4**	**49.6**	**93.3**

Table 5. Comparison with other methods on 3DPW test set. The units for mean joint and vertex errors are in *mm*. † denotes our implementation results.

and mesh generation, the increased time is negligible and has little impact on the final inference time. Accelerating the sequentially estimation can be left for future research.

Visualization. We also give qualitative comparison between relative rotation estimation and SGRE in Fig. 5. Benefited by the powerful neural network and large scale 3D body mesh datasets, both methods can generate natural and well-aligned mesh for images in the wild. However, we can observe that SGRE can generate better mesh, especially at the end of kinematics chains like elbow and knee.

4.4. Comparison with Other Methods

We compare SGRE with previous methods on 3D human mesh recovery benchmark 3DPW [37] and 2D human pose estimation benchmark COCO [28] and CrowdPose [26]. The results are shown in Table 5, 6, 7 and 8.

Table 5 compares SGRE with recent works on 3DPW test set, which is the most widely adopted evaluation benchmark. We compare previous methods that adopt temporal information: HMMR [18] and VIBE [21], multi-stage methods including Pose2Mesh [6] and I2L-MeshNet [32], and single-stage methods: HMR [18], SPIN [23], PARE [22] and 3DCrowdNet [7]. Compared with these methods, the proposed SGRE achieves competitive or superior results. For example, SGRE with 3DCrowdNet achieves 78.4 MPJPE on 3DPW test set, which is lower than PARE by 4.5 mm, 3DCrowdNet by 3.3 mm. It also reduces the PVE to 93.3 mm, outperforming previous methods by a large margin. Note that SGRE is general and can be flexibly integrated into other methods. In Table 5 we show that SGRE can constantly improve the performance of baseline and sota regression-based methods.

Method	3DPW-OCC		
	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Zhang *et al*. [43]	-	72.2	-
SPIN [23]	95.6	60.8	121.6
HMR-EFT [17]	94.4	60.9	111.3
PARE [22]	90.5	56.6	107.9
Baseline	93.7	58.2	110.1
+SGRE	89.9	56.9	104.4
3DCrowdNet†	87.6	54.2	101.2
+SGRE	**83.6**	**53.1**	**97.2**

Table 6. Comparison with other methods on 3DPW-OCC. The units for mean joint and vertex errors are in *mm*. † denotes our implementation results.

Method	COCO				
	AP ↑	AP^{50}	AP^{75}	AP^{M}	AP^{L}
OpenPose [5]	65.3	85.2	71.3	62.2	70.7
HMR [18]	18.9	47.5	11.7	21.5	17.0
GraphHMR [24]	9.3	26.9	4.2	11.3	8.1
SPIN [23]	17.3	39.1	13.5	19.0	16.6
PyMAF [42]	24.6	48.9	22.7	26.0	24.2
Baseline	42.5	74.6	43.5	46.1	40.3
+SGRE	44.6	75.5	46.6	47.6	42.7
3DCrowdNet†	55.4	85.1	64.3	61.8	53.1
+SGRE	**58.4**	**86.5**	**66.2**	**62.5**	**55.6**

Table 7. Comparison with other methods on COCO val set. † denotes our implementation results.

Method	CrowdPose					
	AP ↑	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
SPIN [23]	17.2	42.4	11.2	-	-	-
PyMAF [42]	17.4	42.7	13.0	-	-	-
ROMP [35]	28.5	58.8	24.7	-	-	-
OCHMR [19]	23.6	51.1	18.7	-	-	-
Baseline	41.9	68.9	43.7	51.0	43.1	32.0
+SGRE	43.1	69.3	46.2	52.6	44.2	33.1
3DCrowdNet†	48.0	77.7	51.5	54.2	48.9	41.3
+SGRE	**52.9**	**79.8**	**57.5**	**59.3**	**53.6**	**46.1**

Table 8. Comparison with other methods on CrowdPose test set. † denotes our implementation results.

We also test the proposed SGRE in more challenging scenarios, *i.e.*, occlusion. Following previous work [22], we conduct experiments on 3DPW-OCC, which is a occlusion-specific dataset to evaluate the robustness of methods under occlusion. Table 6 compares SGRE with previous methods, including SPIN [23], HMR-EFT [17] and PARE [22].

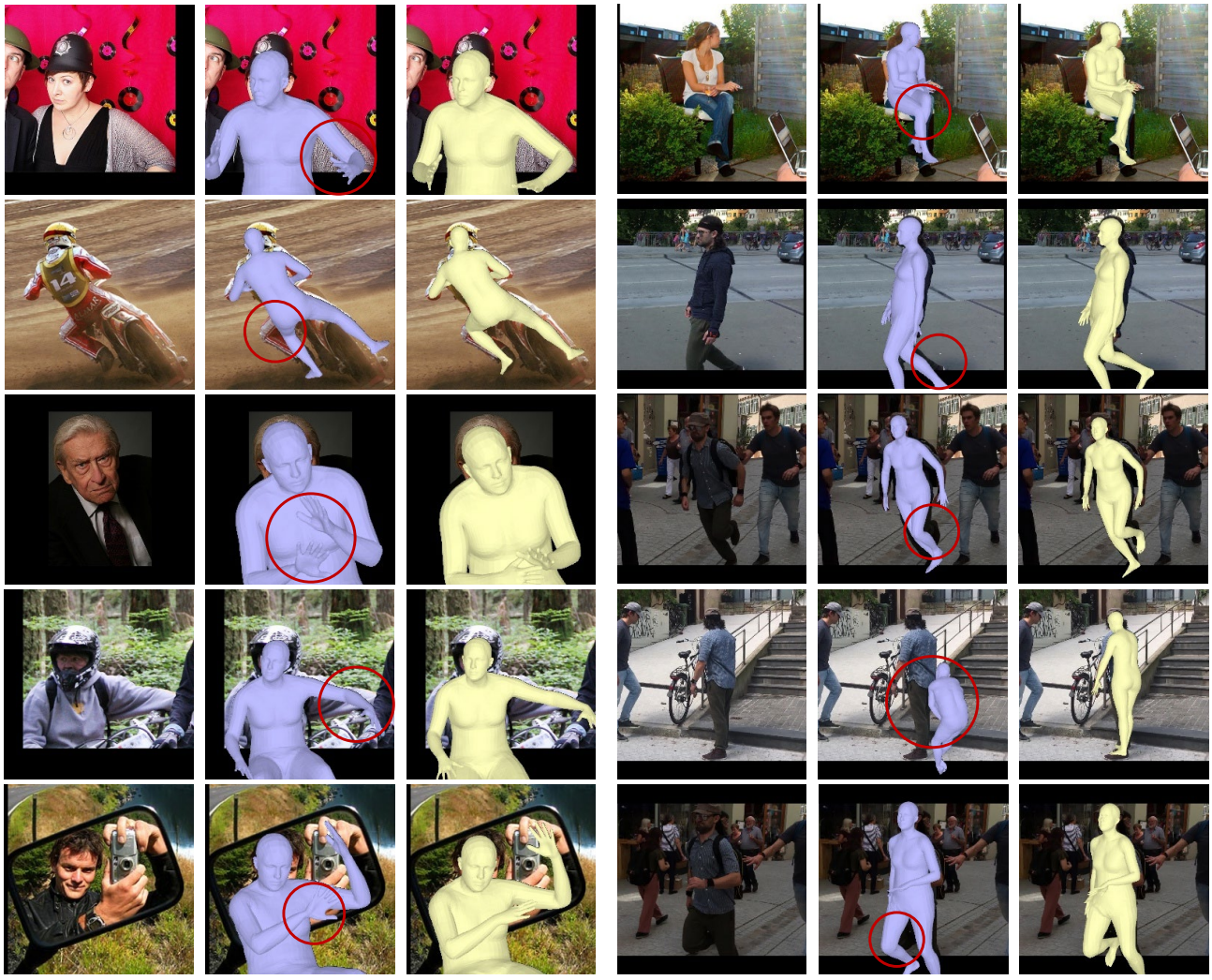


Figure 5. Qualitative results on COCO (columns 1-3) and 3DPW (columns 4-6) datasets. From left to right: (a) Input image, (b) Relative rotation estimation by 3DCrowdNet [7], (c) and our SGRE results.

SGRE with sota 3DCrowdNet achieves 83.6 MPJPE and 97.2 PVE on 3DPW-OCC, outperforming previous methods by a large margin. Table 6 and Table 5 shows that SGRE gets more substantial adavantages on occlusion scenarios. This indicates that the proposed SGRE is better than relative rotation estimation, because it alleviates the error accumulation along the kinematics chains.

Finally, we evaluate 2D human pose estimation performance on COCO val set and CrowdPose test set to verify the effectiveness of SGRE in real-world scenarios. We project the 3D joints regressed from mesh on the image plane and report the performance. The results are shown in Table 7 and Table 8. We can observe that SGRE can significantly improve the performance of both baseline and 3DCrowdNet and achieves superior performance over previous regression-based methods.

5. Conclusion

This paper proposes the SGRE to estimate the accurate global rotations for 3D human mesh recovery. Different from previous methods that estimate relative rotations, the proposed SGRE alleviates the error accumulation along the kinematics chains, leading to smaller joint and mesh vertex error. SGRE successively estimates the global rotation of each joint based on previous estimated results, relieving the difficulty of directly estimation. Experiments on several 3D human mesh recovery and 2D human pose estimation benchmarks demonstrate the effectiveness of SGRE.

Acknowledgement: This work is supported in part by Natural Science Foundation of China under Grant No. U20B2052, 61936011, in part by The National Key Research and Development Program of China under Grant No. 2018YFE0118400.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *SIGGRAPH*. 2005. 1

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 3

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 1, 2

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 7

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2, 3, 7

[7] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[8] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. *NeurIPS*, 32, 2019. 7

[9] Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *CVPR*, 2023. 3

[10] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 5

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 2

[12] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 6

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2013. 3, 5

[15] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *3DV*, 2021. 1, 2, 3, 5

[16] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 2

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

[18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 5, 7

[19] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *CVPR*, 2022. 5, 7

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3, 7

[22] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 1, 2, 3, 5, 7

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

[24] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3, 7

[25] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 1, 2

[26] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 5, 7

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 1, 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 5, 7

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 1, 2, 3

[30] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 7

[31] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 5

[32] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 3, 7

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

[34] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 5

[35] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 7

[36] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 7

[37] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 3, 5, 7

[38] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *CVPR*, 2022. 1

[39] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for instance-level human analysis. *IEEE TPAMI*, 2023. 1

[40] Dongkai Wang, Shiliang Zhang, and Gang Hua. Robust pose estimation in crowded scenes with direct pose-level inference. *NeurIPS*, 2021. 1

[41] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 2

[42] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 1, 3, 5, 7

[43] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 7

[44] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *ICCV*, 2021. 2

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4