SKEL-CF: Coarse-to-Fine Biomechanical Skeleton and Surface Mesh Recovery

Da Li[1,2*], Jiping Jin[1,3*], Xuanlong Yu[1], Wei Liu[5], Xiaodong Cun[4], Kai Chen[5], Rui Fan[3], Jiangang Kong[5], Xi Shen[1,†]

[*]Equal contribution [†]Corresponding author

[1]Intellindust AI Lab [2]Shenzhen University [3]ShanghaiTech University
[4]GVC Lab, Great Bay University [5]Didi Chuxing Co.Ltd

Abstract. Parametric 3D human models such as SMPL have driven significant advances in human pose and shape estimation, yet their simplified kinematics limit biomechanical realism. The recently proposed SKEL model addresses this limitation by re-rigging SMPL with an anatomically accurate skeleton. However, estimating SKEL parameters directly remains challenging due to limited training data, perspective ambiguities, and the inherent complexity of human articulation. In this work, we propose SKEL-CF, a new framework for estimating SKEL parameters. SKEL-CF adopts a standard transformer-based encoder–decoder architecture. The encoder first produces coarse predictions of the camera extrinsics and SKEL parameters. The decoder then iteratively refines these predictions across multiple layers, with explicit and auxiliary supervision applied at each layer. To provide anatomically consistent training data, we convert the existing SMPL-based dataset into a SKEL-aligned version, called HMR-SKEL. This new dataset offers high-quality supervision for SKEL estimation. In addition, to reduce depth and scale ambiguity, we explicitly incorporate camera intrinsic estimation into the SKEL-CF pipeline and show that it is important for accurate reconstruction. Extensive experiments validate the effectiveness of the proposed design. On the challenging MOYO dataset, SKEL-CF achieves 85.0 MPJPE / 51.4 PA-MPJPE, significantly outperforming the previous SKEL-based state-of-the-art HSMR (104.5 / 79.6). These results establish SKEL-CF as a promising framework for human motion analysis, facilitating the use of computer vision techniques in biomechanics-related analysis. Our implementation is available on the project page: `https://pokerman8.github.io/SKEL-CF/`.

1 Introduction

In recent years, 3D human pose and shape estimation has made remarkable progress, enabling a wide range of downstream applications [5,8,19,32,37]. However, adoption in biomechanics, a domain where such techniques could be particularly impactful, remains limited. This gap arises because existing parametric models, such as SMPL [23], SMPL-X [28], and GHUM [40], fall short of the stringent biomechanical requirements. Their simplified kinematics and loosely

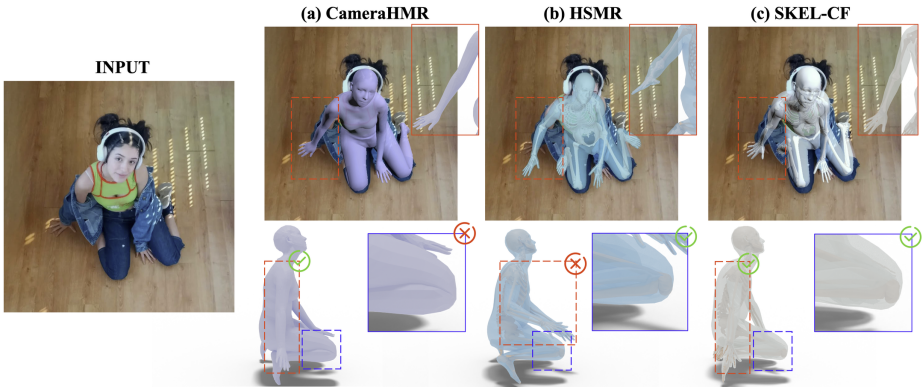

Fig. 1: Visualization of SKEL-CF on web images. Compared with SMPL-based state-of-the-art CameraHMR [27], our SKEL-based model produces more natural joint motions (see the side-view zoomed knee in blue). Compared with HSMR [39], the state-of-the-art SKEL-based method, SKEL-CF achieves more accurate skeleton and mesh reconstruction (see zoomed hand in orange).

constrained axis–angle representations often compromise biomechanical realism, especially in complex articulations such as deep squats or yoga poses [39]. In these scenarios, the estimated parameters must correspond to anatomically accurate skeletons, respect joint limits, and ensure physically plausible motion with high precision. As illustrated in Fig. 1 (a), even the state-of-the-art SMPL-based method CameraHMR [27] can produce anatomically implausible results, such as unnatural knee bending.

To address these limitations, the recently proposed SKEL [14] model redefines the foundation of human representation by re-rigging SMPL with an anatomically accurate skeleton and realistic joint constraints. This makes it a promising step toward bridging computer vision and biomechanics, offering a representation that is both visually coherent and biomechanically faithful. HSMR [39] is the first transformer-based method to reconstruct a human body in the SKEL space from a single image. It converts the predicted SMPL ground truth from existing datasets into SKEL pseudo ground truth for supervised training. However, estimating SKEL from a single image remains challenging because it requires recovering a more constrained 3D structure than SMPL under the inherent depth ambiguity of monocular input. Consequently, the predictions often lack accuracy (see Fig. 1 (b)), and empirical metrics such as MPJPE remain inferior to SMPL-based models.

In this work, we advance SKEL-based human mesh recovery by proposing SKEL-CF, a new framework for estimating SKEL parameters. Similar to HSMR [39], SKEL-CF follows a standard transformer-based encoder–decoder architecture. However, unlike HSMR, which applies supervision only at the final layer, we adopt a coarse-to-fine estimation strategy, which is inspired by DETR [3, 42]: the encoder predicts initial coarse camera extrinsics and SKEL parameters. The decoder then predicts the residual to refine the prediction.

Moreover, by iteratively refining the predictions layer by layer in the decoder, with supervision applied at each decoder layer, we empirically observe further improvement. To handle diverse camera viewpoints, we incorporate camera intrinsic estimation following CameraHMR [27]. This improves robustness across different perspectives and reduces depth and scale ambiguity. Finally, leveraging the high-quality SMPL estimations from CameraHMR [27], we generate large-scale and high-fidelity SKEL annotations to construct HMR-SKEL dataset. This dataset provides reliable supervision and enables more effective training for SKEL estimation. We position SKEL-CF as a SKEL-specific recovery pipeline that jointly addresses annotation quality, perspective ambiguity, and constrained biomechanical parameter regression.

Empirically, SKEL-CF achieves substantial improvements across multiple benchmarks. On the challenging MOYO dataset [33], it achieves 85.0 MPJPE and 51.4 PA-MPJPE, significantly outperforming the previous SKEL-based state-of-the-art HSMR [39], which reports 104.5 and 79.6, respectively. To further evaluate performance on challenging motions, we construct *MOYO-HARD* by removing the first and last 25% of frames from each sequence, where subjects are typically in static poses. On this subset, SKEL-CF achieves 90.0 MPJPE and 61.5 PA-MPJPE, compared with 120.0 and 97.7 for HSMR. In addition, SKEL-CF achieves comparable or better performance than leading SMPL-based methods such as CameraHMR [27] on 3DPW [35], EMDB [13], and MOYO [33], while producing more biomechanically realistic and visually consistent human meshes (Fig 1(c)). We also perform extensive ablation studies to validate the effectiveness of each component in SKEL-CF.

To conclude, our key contributions are summarized as follows:

- We present SKEL-CF as an integrated SKEL-based recovery framework that combines high-quality SKEL supervision, perspective-aware camera conditioning, and coarse-to-fine SKEL regression for faithful mesh and skeleton recovery.
- We introduce HMR-SKEL, a large-scale, high-quality dataset with anatomically consistent SKEL annotations converted from CameraHMR-refined SMPL annotations [27]. This dataset provides reliable supervision and unified skeleton–mesh alignment, enabling stronger SKEL-based model training.
- Our ablation studies demonstrate that camera intrinsic estimation reduces perspective and scale ambiguity, coarse-to-fine estimation (encoder prediction followed by decoder refinement) improves performance, and iterative refinement with layer-wise decoder supervision further enhances robustness, particularly on challenging articulated motions such as MOYO-HARD.

Our implementation is available on the project page: `https://pokerman8.github.io/SKEL-CF/`.

2 Related Work

Parametric human body model. Parametric human body models are widely used for 3D human shape and pose estimation, as they offer a structured low-

dimensional representation of human geometry. SMPL [23] represents the human mesh with fixed topology, using low-dimensional shape and pose parameters with linear blend skinning and pose-dependent corrective shapes for realistic articulation. SMPL-H [29] augments this framework with articulated hands, while SMPL-X [28] further integrates the FLAME [20] facial model, yielding a unified representation of body, hands, and expressive face. While the SMPL family provides a compact and effective representation of the human body surface, it employs a simplified kinematic structure that deviates from the anatomical skeletal system of real humans, allowing anatomically implausible poses (e.g., the knee can exhibit unrealistically free rotation, see Fig. 1 (b)). SKEL [14] introduces anatomically accurate joint definitions and a bone hierarchy embedded within a parametric human body model, overcoming limitations of simplified kinematic structures in models such as SMPL. By estimating parameters that are explicitly compatible with biomechanical skeletons, SKEL enforces realistic joint limits and produces physically plausible motion, leading to more accurate and reliable pose reconstruction.

Human mesh recovery. The lack of 3D annotations for in-the-wild images remains a major obstacle in Human mesh recovery. SMPLify [2] mitigates this by iteratively optimizing SMPL [23] parameters from 2D keypoints, yielding accurate results but at high computational cost. HMR [10] introduces the first end-to-end framework to predict 3D pose and shape from in-the-wild images using only 2D supervision, where adversarial learning with mocap data provides strong pose and shape priors to compensate for the missing 3D information. 4DHuman [6] constructs a large-scale dataset by fitting approximately 3 million images with ProHMR [17], generating pseudo-ground-truth (p-GT) 3D annotations alongside 2D keypoints, and further upgrades the network architecture from CNN to Vision Transformer (ViT). However, TokenHMR [4] identifies a mismatch between 3D and 2D keypoints, where enforcing 2D keypoint alignment can degrade 3D evaluation metrics, partly due to the use of fixed camera intrinsics and p-GT annotations in the 4DHuman [6] dataset. To address this, TokenHMR [4] introduces the TALS loss, which down-weights the 2D keypoint loss when the L2 distance between predicted and ground-truth keypoints falls within a predefined threshold, thus preventing overfitting to noisy p-GT annotations. In addition, it leverages a quantized token dictionary, constructed by pre-training a Vector Quantized-VAE (VQ-VAE) [34] on motion capture datasets [25, 33], to mitigate 3D pose ambiguity. CameraHMR [27] improves the 4DHuman [6] dataset through CamSMPLify, which refines body representations using dense keypoints to overcome the average body limitation. To tackle the fixed-camera constraint, it introduces HumanFOV [27], transforming weak-perspective projection into a fully perspective projection.

SMPL-to-SKEL dataset conversion. The first large-scale 3D pseudo-ground-truth (p-GT) dataset was introduced by 4DHuman [6], which extends to unlabeled datasets such as InstaVariety [11], AVA [7], and AI Challenger [31] by generating p-GT annotations. For each image, an off-the-shelf detector [3] and a body keypoint estimator are applied to obtain bounding boxes and 2D key-

points. A SMPL [23] mesh is then fitted to these keypoints using ProHMR [17], producing pseudo-ground-truth SMPL parameters. Despite its scale and utility, the 4DHuman dataset still contains inaccuracies and artifacts. To that end, CameraHMR [27] refines the dataset using its proposed CamSMPLify method, alongside additional techniques to enhance annotation quality. While large-scale datasets of SMPL parameters have been constructed, there remains a need to develop datasets for SKEL [14] parameters. HSMR [39] addresses this by first converting the 4DHuman SMPL parameters to SKEL parameters using SKEL fitting [14]. To mitigate the inaccuracies in 4DHuman annotations, we further apply SKEL fitting to the refined SMPL annotations provided by CameraHMR [27].

3 Method

In this section, we introduce our approach, SKEL-CF, whose overall pipeline is illustrated in Fig. 2. SKEL-CF aims to estimate SKEL parameters from a single image using an encoder–decoder architecture that performs coarse-to-fine estimation. The encoder produces initial predictions of the camera extrinsics $\boldsymbol{\pi}_0$, shape parameters $\boldsymbol{\beta}_0$, and pose parameters $\boldsymbol{\theta}_0$, while the decoder iteratively refines these predictions across layers. To enhance robustness to camera variations, we adopt the pretrained camera intrinsic predictor from CameraHMR [27] to estimate the focal length, keeping its parameters frozen throughout training. This section is organized as follows. We first introduce the SKEL model and the HMR-SKEL dataset that we construct to supervise SKEL learning in Section 3.1. We then describe the details of SKEL-CF and its implementation in Section 3.2.

3.1 Preliminaries

SKEL model. The SKEL model [14] is a parametric representation of the human body that enables the unified reconstruction of both the surface mesh and the skeleton mesh. Given the pose parameters $q \in \mathbb{R}^{46}$ and the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ (corresponding to the top 10 shape components), SKEL generates the human mesh $M \in \mathbb{R}^{6890\times3}$ and the skeleton mesh $S_k \in \mathbb{R}^{24752\times3}$. Unlike SMPL, which represents pose parameters $\theta \in \mathbb{R}^{72}$ using unconstrained three-degree-of-freedom ball joints, SKEL constrains the motion of each joint. This reduces the pose parameter space from 72 to 46. For example, SKEL models the elbow as a hinge joint by limiting its degrees of freedom. This formulation ensures anatomically consistent and biomechanically valid poses [14, 39]. Consequently, SKEL produces physically plausible human motion and is particularly suitable for downstream applications such as motion analysis, rehabilitation, and human–robot interaction.

HMR-SKEL dataset. HSMR [39] generated pseudo ground-truth (p-GT) SKEL parameters by converting SMPL parameters into the SKEL space. While this provided a practical starting point, its quality was limited by low-resolution images and noisy SMPL annotations in the original dataset. With the release

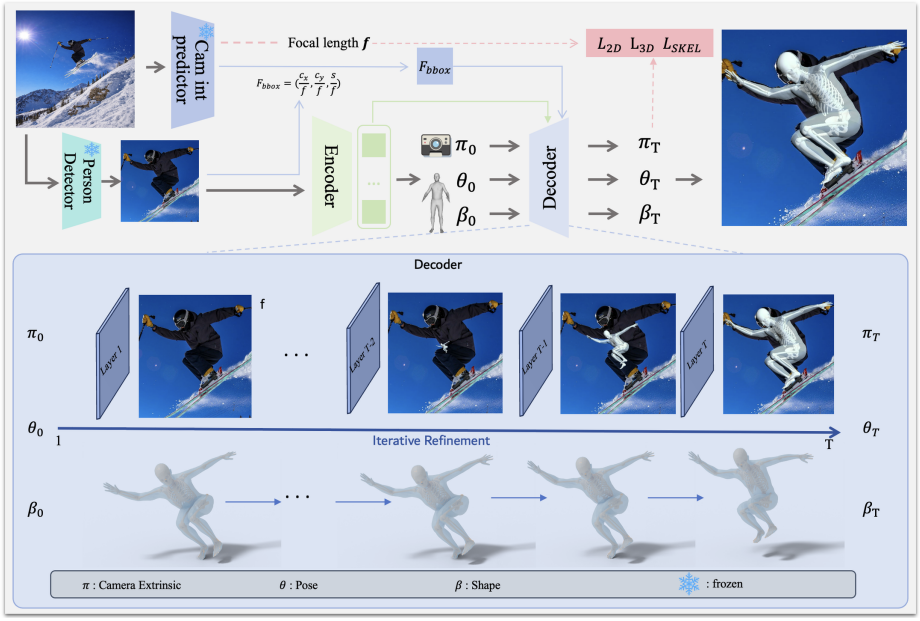

Fig. 2: Overview of the proposed SKEL-CF. Our method estimates SKEL parameters from a single image using an encoder–decoder architecture that performs coarse-to-fine estimation. The encoder produces initial predictions of the *camera extrinsics* $\boldsymbol{\pi}$, *shape parameters* $\boldsymbol{\beta}$, and *pose parameters* $\boldsymbol{\theta}$. The decoder then refines these predictions iteratively across layers. In addition, we adopt the pretrained camera model from CameraHMR [27] to estimate the focal length, and keep its parameters frozen throughout training.

of refined SMPL annotations from CameraHMR [27], we revisit this conversion process to obtain higher-quality SKEL pseudo ground truth. Following the HSMR setup, which is based on HMR2.0 [6], we use data from Human3.6M [9], MPI-INF-3DHP [26], COCO [22], MPII [1], AI Challenger [38], and InstaVariety [12]. We exclude AVA because CameraHMR does not provide refined SMPL annotations for this subset. We denote the resulting dataset as HMR-SKEL.

Our optimization pipeline follows the fitting protocol of SKEL [14]. We first reconstruct a reference human mesh from the high-quality annotations provided by CameraHMR [27], which is used as the target SMPL mesh. Given an initial set of SKEL parameters, we treat them as learnable variables and feed them into the SKEL model to generate a SKEL mesh. We then compute the alignment loss between the generated SKEL mesh and the target SMPL mesh, and use this loss to iteratively update the SKEL parameters. To improve optimization stability, we perform the fitting in a hierarchical manner: we first optimize the lower-body parameters, then the upper-body parameters, and finally the full-body parameters. Moreover, since both SMPL and SKEL share the same global orientation and translation terms, we keep these variables fixed during optimization; em-

pirically, allowing them to vary leads to worse alignment. The full process takes about 58 hours to process 3M images on a single RTX 3090 GPU.

3.2 Our method: SKEL-CF

Overall pipeline. As illustrated in Fig. 2, SKEL-CF adopts an encoder–decoder architecture with a coarse-to-fine strategy to estimate SKEL parameters from a single RGB image. The target parameter set is defined as $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\pi}\}$, where $\boldsymbol{\theta}$ denotes human pose, $\boldsymbol{\beta}$ is body shape, and $\boldsymbol{\pi}$ indicates camera extrinsics. Given an input image, following previous work [27], we first detect the person using an off-the-shelf human detector and convert the cropped region into visual tokens. We also estimate the camera intrinsic parameter, specifically the focal length f, from the full image using a pretrained predictor from CameraHMR [27], which remains frozen during training. The visual tokens are passed through the encoder to produce initial predictions $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \boldsymbol{\pi}_0\}$ along with contextual visual features. These predictions are then progressively refined through T decoder layers. At each stage, refinement is guided by the encoded visual features and geometric cues $\mathcal{F}_{\text{bbox}} = \left(\frac{c_x}{f}, \frac{c_y}{f}, \frac{s}{f}\right)$, where (c_x, c_y) and s denote the bounding box center and scale. After T refinement steps, the final output is $\boldsymbol{\Theta}_T = \{\boldsymbol{\theta}_T, \boldsymbol{\beta}_T, \boldsymbol{\pi}_T\}$, yielding progressively improved SKEL parameter estimates. The focal-normalized feature $\mathcal{F}_{\text{bbox}}$ is injected into the decoder as bounding-box geometry, where it guides projection-sensitive residual corrections after the image crop has been encoded.

Learning objectives. Denoting the target parameters as $\hat{\boldsymbol{\Theta}} = \{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}\}$, the learning objective integrates keypoint-level and parameter-level supervision:

$$\mathcal{L}_{\text{skel}}(\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}) = \underbrace{\lambda_{kp}\mathcal{L}_{\text{kp}}}_{\text{keypoint-level}} + \underbrace{\lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{\theta}\mathcal{L}_{\theta}}_{\text{parameter-level}}. \tag{1}$$

where λ_{kp}, λ_{β}, and λ_{θ} are hyperparameters controlling the relative weight of each term. This formulation allows us to simultaneously supervise the reconstructed 3D/2D joints and the underlying pose and shape parameters. Note that no direct camera extrinsic supervision $\hat{\boldsymbol{\pi}}$ is provided; the ability to predict camera extrinsics is only implicitly learned through 2D keypoint reprojection, making the training challenging.

Keypoint-level supervision. Given the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$, we first reconstruct 3D joint positions $\mathbf{J}_{3d} \in \mathbb{R}^{K\times 3}$ via forward kinematics, where K is the number of joints. These 3D joints are projected onto the image plane using the predicted camera parameters to obtain 2D joint locations $\mathbf{J}_{2d} \in \mathbb{R}^{K\times 2}$. The coarse keypoint loss $\mathcal{L}_{\text{kp}}$ enforces consistency with pseudo-ground-truth 3D and 2D joints:

$$\mathcal{L}_{\text{kp}} = \|\hat{\mathbf{J}}_{3d} - \mathbf{J}_{3d}\|_1 + \|\hat{\mathbf{J}}_{2d} - \mathbf{J}_{2d}\|_1. \tag{2}$$

Parameter-level supervision. Since the proposed dataset HMR-SKEL provides reliable pose $\hat{\boldsymbol{\theta}}$ and shape $\hat{\boldsymbol{\beta}}$ parameters, we impose supervision directly on these values:

$$\mathcal{L}_{\beta} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1, \quad \mathcal{L}_{\theta} = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1. \tag{3}$$

Coarse-to-fine refinement. SKEL-CF adopts a coarse-to-fine refinement strategy. The encoder first produces an initial estimate of the parameters, denoted as $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \boldsymbol{\pi}_0\}$. The decoder then refines these predictions across T layers, where the i-th layer outputs $\boldsymbol{\Theta}_i = \{\boldsymbol{\theta}_i, \boldsymbol{\beta}_i, \boldsymbol{\pi}_i\}$. The final layer produces the refined prediction $\boldsymbol{\Theta}_T$. The encoder prediction and the final decoder prediction are supervised using the same objective function $\mathcal{L}_{\text{skel}}$:

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{skel}}(\boldsymbol{\Theta}_0, \hat{\boldsymbol{\Theta}}), \quad \mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{skel}}(\boldsymbol{\Theta}_T, \hat{\boldsymbol{\Theta}}) \tag{4}$$

Iterative refinement. We use iterative refinement with layer-wise supervision. The refinement process consists of two operations: coarse-to-fine initialization and residual refinement. The encoder first predicts an image-dependent coarse estimate $\boldsymbol{\Theta}_0$, and the decoder then updates this estimate *layer by layer* by predicting residual corrections. To supervise the intermediate decoder layers efficiently, we apply an auxiliary pose-parameter loss:

$$\mathcal{L}_{\text{refine}} = \sum_{i=1}^{T-1} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1, \tag{5}$$

where T denotes the total number of decoder layers, and $\hat{\boldsymbol{\theta}}_i$ denotes the corresponding ground-truth pose parameters. We observe that optimizing $\mathcal{L}_{\text{skel}}$ at intermediate decoder layers achieves performance comparable to $\mathcal{L}_{\text{refine}}$. However, as computing keypoint-level predictions through the SKEL forward process is computationally expensive, we limit the optimization to the pose parameters for efficiency.

Note that coarse-to-fine strategy and iterative refinement are widely adopted in the DETR-based object detection frameworks [3, 42]. The experiments to demonstrate the effectiveness of iterative refinement are provided in the supplementary material.

Overall optimization goal. The total training objective $\mathcal{L}_{\text{tot}}$ aggregates the encoder and decoder losses, together with the refinement term:

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{dec}} + \mathcal{L}_{\text{enc}} + \lambda_{ref}\mathcal{L}_{\text{refine}}, \tag{6}$$

where λ_{ref} is the coefficient for the refinement loss.

Implementation details. Following HSMR [39], we adopt ViTPose-H [41], pretrained on COCO [22], as the encoder backbone. ViTPose-H consists of 32 transformer layers, each with 16 attention heads and a hidden dimension of 1280. The decoder is a standard transformer decoder with $L = 6$ layers, consistent with the design in HSMR [39]. We train SKEL-CF using the AdamW [24] optimizer with β_1, β_2 setting to 0.9, 0.999, weight decay setting to 1×10^{-4}, a batch size of 64 and a learning rate of 1×10^{-5}, preceded by a one-epoch warm-up. Training is conducted for 30 epochs on 8 NVIDIA A100 GPUs, taking approximately 120 hours in total. The hyper-parameters are set as $\lambda_{kp} = 0.05$, $\lambda_{\beta} = 0.0005$, $\lambda_{\theta} = 0.001$, and $\lambda_{ref} = 0.1$. Notably, our training schedule is significantly shorter than that of HSMR [39], which trains for 100 epochs.

4 Experiment

In this section, we first introduce the datasets and evaluation metrics in Section 4.1. We then provide a comparison between SKEL-based methods in Section 4.2 and SMPL-based approaches in Section 4.3. Finally, we present the ablation studies and additional discussions in Section 4.4.

4.1 Datasets and evaluation metrics

Evaluated datasets. We evaluate SKEL-CF on five representative datasets covering diverse environments and motion complexities, following standard practice [4, 10, 27, 39]. *3DPW* [35] provides in-the-wild videos with accurate 3D annotations captured by moving cameras. *Human3.6M* [9] includes controlled indoor actions (e.g., sitting, walking, greeting) with motion-capture ground truth. *EMDB* [13] offers high-quality pose and shape sequences captured with electromagnetic sensors and handheld cameras. *SPEC-SYN* [16] is a synthetic dataset with diverse camera variations, enabling controlled evaluation under varying camera settings. In contrast, *MOYO* [33] contains large-scale yoga sequences with extreme poses, frequent self-occlusions, and ground contact, providing the most challenging and diverse test scenario.

Since most yoga videos in MOYO starts and ends with the same static poses, we select the front-view camera among the eight available viewpoints and remove the first and last 25% of frames to retain the complex and diverse motion segments, forming the curated *MOYO-HARD* subset. More details about MOYO-HARD are provided in the supplementary material, where we include visualizations of the removed first and last 25% segments of the MOYO videos.

Evaluation metrics. We evaluate the 3D human body recovery using MPJPE, PA-MPJPE, and PVE. Among them, MPJPE and PA-MPJPE serve as sparse evaluations, measuring the Euclidean distance between the predicted and ground-truth kinematic joints before and after rigid alignment, respectively. In contrast, PVE provides a dense evaluation, computing the vertex-level error between the predicted and ground-truth meshes, thereby offering a more comprehensive assessment of surface reconstruction accuracy. The 2D keypoint metrics, PCK@0.05 and PCK@0.1, which quantify the ratio of projected keypoints within a normalized distance threshold to the ground truth.

4.2 Comparison with SKEL-based approaches.

Quantitative results. We compare SKEL-based methods in Table 1 and Table 2. Among end-to-end approaches, SKEL-CF consistently outperforms prior work HSMR [39] across all datasets. On the challenging MOYO [33] benchmark, SKEL-CF achieves up to 18.6% MPJPE and 35.4% PA-MPJPE improvements, demonstrating strong robustness under large pose variations and occlusions. We also evaluate two two-stage baselines that first estimate SMPL parameters and then fit them to SKEL: HMR2.0 + SKEL fitting and CameraHMR + SKEL

Fig. 3: Visual comparison between the proposed SKEL-CF and HSMR [39]. Our proposed SKEL-CF achieves more precise skeletal estimations (best viewed in zoom). Additional visual results are provided in the supplementary material.

Table 1: Quantitative comparison of skeleton and surface mesh recovery using the SKEL [14] human model on 3DPW [35], Human3.6M [9], and MOYO [33]. The proposed SKEL-CF delivers consistent accuracy gains over previous end-to-end approach. * denotes results reported by [39], ◇ denotes fitting the outputs of [27] to the SKEL [14] model.

Methods	3DPW [35]		Human3.6M [9]		MOYO [33]		MOYO-HARD	
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓
Two-stage Approaches								
HMR2.0 + SKEL fit [6, 39]*	81.0	54.4	53.6	34.1	130.5	93.7	-	-
CameraHMR + SKEL fit [27, 39]◇	70.4	41.8	-	-	**75.5**	**49.9**	**88.7**	**61.4**
End-to-end Approaches								
HSMR [39]	81.5	54.8	50.4	32.9	104.5	79.6	120.0	97.7
SKEL-CF (Ours)	**61.5**	**38.7**	**39.0**	**31.2**	85.0	51.4	90.0	61.5

Table 2: Quantitative comparison of mesh recovery using the SMPL [23] and SKEL [15] human model on 3DPW [35], EMDB [13], and SPEC-SYN [16].The complete comparison including MOYO and MOYO-HARD is provided in Table **??** of the supplementary material.

Method	3DPW [35]			EMDB [13]			SPEC-SYN [16]		
	MPJPE↓	PA-MPJPE↓	PVE↓	MPJPE↓	PA-MPJPE↓	PVE↓	MPJPE↓	PA-MPJPE↓	PVE↓
SMPL-based Approaches									
SPEC [16]	96.5	53.2	118.5	138.9	87.7	161.3	83.5	56.9	98.9
CLIFF [21]	69.0	43.0	81.2	103.5	68.3	123.7	128.5	55.8	139.0
HMR2.0a [6]	69.8	44.4	82.2	97.8	61.5	120.0	133.3	55.8	153.0
TokenHMR [4]	70.5	43.8	86.0	88.1	49.8	104.2	110.5	51.8	127.6
WHAM [30]	57.8	35.9	68.7	79.7	50.4	94.4	-	-	-
ReFit [36]	57.6	38.2	67.6	91.7	55.5	106.2	103.6	51.3	116.3
CLIFF [21]	72.0	46.6	85.0	97.1	61.3	113.2	109.9	55.6	124.6
HMR2.0b [6]	81.3	54.3	93.1	118.5	79.2	140.6	150.7	67.6	172.9
CameraHMR [27]	62.7	38.7	**73.4**	73.2	**43.9**	85.6	**66.0**	**37.0**	**79.1**
SKEL-based Approaches									
HSMR [39]	81.5	54.8	-	-	-	-	-	-	-
SKEL-CF (Ours)	**61.5**	**38.7**	73.5	**72.0**	44.5	**84.7**	69.4	37.1	83.4

fitting. While CameraHMR + SKEL fit achieves competitive results on MOYO and MOYO-HARD, its performance drops on 3DPW, indicating limited stability across datasets. Moreover, the fitting-based approach is computationally expensive (approximately 4 minute per image on RTX 3090), whereas SKEL-CF runs in a single forward pass (approximately 0.8 second per image), making it both more efficient and more reliable.

Visual results. We provide a visual comparison with HSMR [39] in Fig. 3. As shown, the proposed SKEL-CF achieves more accurate skeleton and surface reconstruction. More visual examples similar to Fig. 3 are included in supplementary material.

4.3 Comparison with SMPL-based approaches

Quantitative results. We further compare SKEL-CF with recent state-of-the-art SMPL-based human mesh recovery methods across four representative datasets: 3DPW [35], EMDB [13], SPEC-SYN [16], MOYO [33] and more challenging MOYO-HARD, as summarized in Table 2. Note that SMPL [23] is less constrained compared to SKEL [14], enabling strong quantitative performance

Fig. 4: Visual comparison between the proposed SKEL-CF and CameraHMR [27]. The proposed SKEL-CF is built upon the SKEL [14] representation, which enforces anatomically consistent joint motion, resulting in more natural pose predictions compared to the SMPL [23]-based CameraHMR [27] (best viewed in zoom). Additional visual examples are provided in the supplementary material.

Table 3: Ablation study on MOYO-HARD and COCO. *Cam* denotes camera intrinsic modeling, *C2F* denotes coarse-to-fine estimation, and *Refine* denotes iterative refinement.

Method	Components				MOYO-HARD			COCO [22]	
	Cam	C2F	Refine	Dataset	MPJPE↓	PA-MPJPE↓	PVE↓	PCK@0.05↑	PCK@0.1↑
Baseline (HSMR [39])	✗	✗	✗	HMR2.0 + SKELify	120.0	97.7	140.5	**0.86**	**0.96**
Baseline w. HMR-SKEL	✗	✗	✗	HMR-SKEL	103.6	67.4	121.4	0.76	0.91
Ours w.o Cam	✗	✓	✓	HMR-SKEL	98.8	66.1	113.7	0.79	0.93
Ours w.o C2F	✓	✗	✓	HMR-SKEL	91.5	63.1	105.6	0.67	0.91
Ours w.o Refine	✓	✓	✗	HMR-SKEL	92.7	65.4	107.7	0.77	0.92
Only Cam	✓	✗	✗	HMR-SKEL	92.7	66.4	107.4	0.77	0.92
Ours	✓	✓	✓	HMR-SKEL	**90.0**	**61.5**	**102.5**	0.80	0.93

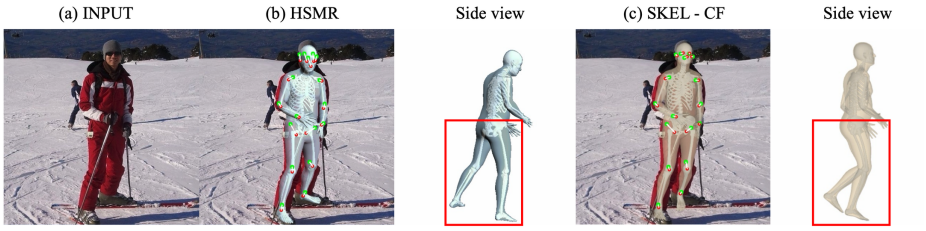

Fig. 5: COCO 2D metric saturation. Similar 2D PCK can hide large 3D pose differences. Red and green denote predicted keypoints and ground-truth 2D keypoints, respectively.

but occasionally producing anatomical inconsistency or unnatural joint motions(see Fig. 1). Despite this inherent difference, SKEL-CF achieves comparable performance to CameraHMR [27], the strongest SMPL-based model trained on 4DHuman dataset, demonstrating that kinematically constrained SKEL representations can reach a similar level of numerical precision while maintaining stronger physical plausibility. Note that, the the 2D keypoint results are provided in the supplementary material.

Visual results. We provide a visual comparison with the SMPL [23]-based state-of-the-art method CameraHMR [27] in Fig. 4. As shown, the proposed SKEL-CF produces anatomically more consistent joint motions, benefiting from the structural constraints of the SKEL [14] model. In contrast, CameraHMR tends to generate unnatural joint configurations under complex motion scenarios. Additional visualizations similar to Fig. 4 are provided in supplementary material.

4.4 Discussion

Ablation study. To analyze the contribution of each component, we conduct controlled ablations on MOYO-HARD, and COCO [22]. Note that more ablation on 3DPW [35], MOYO [33] are provided in the supplementary material. As shown in Table 3. We start from an HSMR-style baseline that disables the camera intrinsic predictor, coarse-to-fine (C2F) initialization, and iterative refinement. Comparing HSMR with the same architecture trained on HMR-SKEL isolates

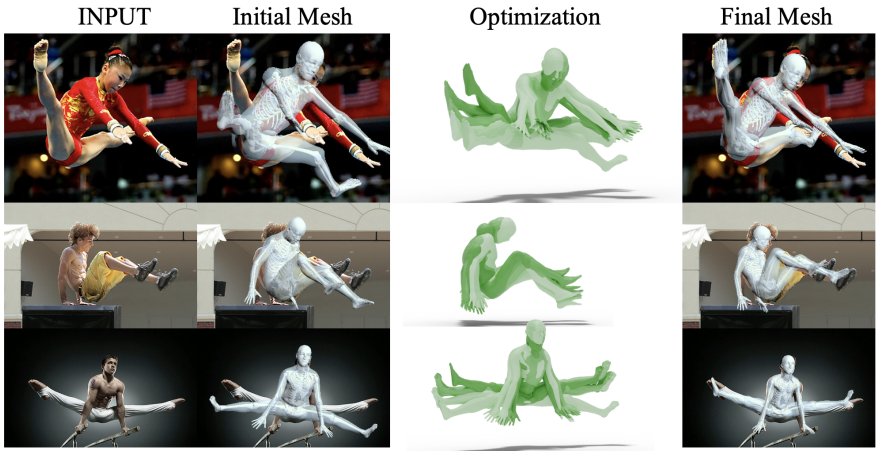

Fig. 6: Illustration of the mesh refinement. The light green meshes represent the initial coarse estimations, while the dark green meshes denote the iteratively refined final results.

the effect of the improved annotations: on MOYO, PA-MPJPE improves from 79.6 to 53.7. Building on this same-data baseline, the full SKEL-CF pipeline further improves MOYO PA-MPJPE from 53.7 to 51.4, showing additional gains beyond the dataset. C2F and refinement bring more visible gains on the harder MOYO-HARD subset, where the full model improves PVE over the Cam-only variant from 107.4 to 102.5, and also recovers stronger COCO 2D alignment than the without C2F or without iterative refinement variants. Note that COCO PCK evaluates only the 2D projection of keypoints and can be insensitive to 3D pose errors. Fig. 5 shows that similar 2D PCK can correspond to noticeably different 3D reconstructions, so we treat COCO PCK as a complementary projection metric rather than the primary evidence for SKEL recovery quality.

Iterative mesh refinement. We visualize the iterative mesh refinement process of SKEL-CF in Fig. 6, which clearly illustrates the effectiveness of the proposed coarse-to-fine strategy. Additional qualitative results are provided in the supplementary material. Quantitative validation is reported in Table 4. We further investigate two variants: *i)* applying the full supervision loss L_{skel} to intermediate decoder layers (*Full super.*), instead of using only the refinement loss L_{refine} as sparse supervision; *ii)* replacing the six-layer decoder with a single decoder layer that is reused six times [18] (*Looped Decoder*). Results show that sparse supervision achieves performance comparable to full supervision. However, computing $\mathcal{L}_{\text{skel}}$ is significantly more expensive, as it requires forwarding the SKEL model to obtain the keypoint loss. Therefore, sparse supervision is more computationally efficient and preferable in practice. In contrast, replacing the multi-layer decoder with a single shared layer substantially reduces the decoder's representational capacity, leading to inferior reconstruction quality and weaker generalization.

Table 4: Ablation study of iterative refinement. *Sparse super.* applies refinement supervision across decoder layers, *Full super.* indicates supervision with full SKEL loss $\mathcal{L}_{skel}$ at every layer, and *Looped Decoder* repeats a single layer for multiple refinement steps. "Decoder Layer × Refine Time" summarizes the refinement configuration.

Method	Loss	Decoder Layer × Refine Time	3DPW [35]			MOYO [33]			MOYO-HARD		
			MPJPE↓	PA↓	PVE↓	MPJPE↓	PA↓	PVE↓	MPJPE↓	PA↓	PVE↓
Sparse super. (Ours)	$\mathcal{L}_{refine}$	6×1	61.5	38.7	73.5	**85.0**	**51.4**	**91.9**	**90.0**	**61.5**	**102.5**
Full super.	$\mathcal{L}_{skel}$	6×1	**61.3**	**38.6**	**73.3**	86.4	53.2	93.9	92.1	65.2	105.6
Looped Decoder	$\mathcal{L}_{refine}$	1×6	62.5	40.1	74.8	86.8	54.4	94.2	96.8	70.1	111.5

Table 5: Impact of the encoder's coarse camera translation on EMDB [13]. We compare the encoder's coarse camera translation against a PnP-based estimate computed from ground-truth 2D keypoint annotations.

Method	L2↓	$t_x \downarrow$	$t_y \downarrow$	$t_z \downarrow$	FPS on H100
Encoder coarse trans.	0.4223	0.0414	0.0294	0.4127	91.7
PnP trans. w. G.T. 2D Keypoints	0.4188	0.0403	0.0296	0.4096	84.7

Impact of the encoder's coarse camera translation. We compare the encoder's coarse camera translation with a PnP-based camera translation on EMDB [13]. As shown in Table 5, PnP provides only marginal improvement over the encoder prediction while slightly reducing the inference speed from 91.7 FPS to 84.7 FPS. Notably, this comparison favors PnP by using ground-truth 2D keypoint annotations. In practical deployments, where ground-truth keypoints are unavailable, PnP would require an additional 2D keypoint detector, introducing further computational overhead. Therefore, we adopt the encoder's single-forward-pass camera prediction in SKEL-CF.

5 Conclusion

In this work, we introduced SKEL-CF, a biomechanical skeleton and surface mesh recovery pipeline built upon the SKEL representation. By transforming the high-quality SMPL dataset into a SKEL-based format, termed HMR-SKEL, we established a strong foundation for accurate and generalizable training. Our integration of an explicit camera intrinsic predictor effectively mitigates the limitations of weak-perspective assumptions seen in previous works such as HSMR, enabling robust pose estimation under diverse viewpoints. Furthermore, the proposed coarse-to-fine refinement strategy enables stable and high-fidelity reconstruction results. Overall, SKEL-CF sets a new baseline for biomechanical skeleton and surface recovery, providing a practical and extensible framework for future research in human modeling and motion understanding.

Acknowledgment We gratefully acknowledge the financial support from Intellindust, especially Dr. Caizhi Zhu and Xiao Zhou. We also thank the DiDi Infra Team and Great Bay University for providing computational resources. Finally, we sincerely thank the anonymous reviewers for their insightful comments and constructive feedback, which greatly helped improve this paper.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV (2016)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
4. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In: CVPR (2024)
5. Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., Finn, C.: Humanplus: Humanoid shadowing and imitation from humans. In: CoRL (2024)
6. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
7. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018)
8. He, T., Luo, Z., Xiao, W., Zhang, C., Kitani, K., Liu, C., Shi, G.: Learning human-to-humanoid real-time whole-body teleoperation. In: IROS (2024)
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI (2013)
10. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
11. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
12. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: CVPR (2019)
13. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In: ICCV (2023)
14. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. TOG (2023)
15. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3d digital humans. TOG (2023)
16. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing people in the wild with an estimated camera. In: ICCV (2021)
17. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV (2021)
18. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
19. Li, J., Zhu, Y., Xie, Y., Jiang, Z., Seo, M., Pavlakos, G., Zhu, Y.: Okami: Teaching humanoid robots manipulation skills through single video imitation. In: CoRL (2024)
20. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. (2017)

21. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: ECCV (2022)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl. TOG (2015)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
26. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved CNN supervision. In: 3DV (2017)
27. Patel, P., Black, M.J.: Camerahmr: Aligning people with perspective. In: 3DV (2025)
28. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
29. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. TOG (2017)
30. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: CVPR (2024)
31. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: ICCV (2019)
32. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2022)
33. Tripathi, S., Müller, L., Huang, C.H.P., Taheri, O., Black, M.J., Tzionas, D.: 3d human pose estimation via intuitive physics. In: CVPR (2023)
34. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS (2017)
35. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
36. Wang, Y., Daniilidis, K.: Refit: Recurrent fitting network for 3d human recovery. In: ICCV (2023)
37. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: CVPR (2022)
38. Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al.: Ai challenger: A large-scale dataset for going deeper in image understanding. arXiv (2017)
39. Xia, Y., Zhou, X., Vouga, E., Huang, Q., Pavlakos, G.: Reconstructing humans with a biomechanically accurate skeleton. In: CVPR (2025)
40. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR (2020)
41. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)
42. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv (2020)