Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot

Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier,
Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*

NAVER LABS Europe
https://github.com/naver/multi-hmr
*Equal contribution

Abstract. We present Multi-HMR, a strong single-shot model for *multi-person* 3D human mesh recovery from a single RGB image. Predictions encompass the *whole body*, *i.e.*, including hands and facial expressions, using the SMPL-X parametric model and *3D location* in the camera coordinate system. Our model detects people by predicting coarse 2D heatmaps of person locations, using features produced by a standard Vision Transformer (ViT) backbone. It then predicts their whole-body pose, shape and 3D location using a new cross-attention module called the Human Prediction Head (HPH), with one query attending to the entire set of features for each detected person. As direct prediction of fine-grained hands and facial poses in a single shot, *i.e.*, without relying on explicit crops around body parts, is hard to learn from existing data, we introduce CUFFS, the Close-Up Frames of Full-body Subjects dataset, containing humans close to the camera with diverse hand poses. We show that incorporating it into the training data further enhances predictions, particularly for hands. Multi-HMR also optionally accounts for *camera intrinsics*, if available, by encoding camera ray directions for each image token. This simple design achieves strong performance on whole-body and body-only benchmarks simultaneously: a ViT-S backbone on 448×448 images already yields a fast and competitive model, while larger models and higher resolutions obtain state-of-the-art results.

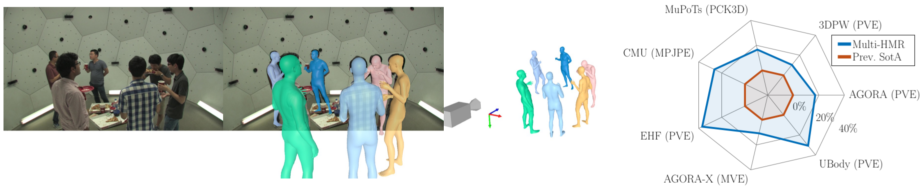


Image Predictions Side View Relative improvement (%)

Fig. 1: Efficient 3D reconstruction of multiple humans in camera space. We introduce Multi-HMR, a single-shot approach to detect *multiple humans* in images, and regress *whole-body* human meshes. Predictions encompass hands and facial expressions, as well as 3D location with respect to the camera. *Left:* Visualization of Multi-HMR predictions. *Right:* Relative improvements (in %) on human mesh recovery benchmarks.

1 Introduction

We introduce a single-shot model for recovering whole-body 3D meshes of humans from a single RGB image. Our problem formulation focuses on four aspects of Human Mesh Recovery (HMR) that we identify as pivotal to making HMR applicable to real-world scenarios: i) capture of expressive body poses – *i.e.*, including hands and facial expressions, ii) efficient processing of images with a variable number of people, iii) location of people in 3D space, iv) adaptability to camera information when available.

Successfully handling these aspects simultaneously makes our proposed model, denoted Multi-HMR, widely applicable. For instance, in virtual or augmented reality (AR/VR), capturing faces and hands precisely is key for communication. It is also beneficial for enabling human-robot interactions [8,49], or human understanding from images and videos [45,50,62]. Likewise, understanding the placement of people in the scene is necessary for applications ranging from robotic navigation to AR/VR applications involving several people. In addition, efficient processing of a variable number of people is desirable when computation is restricted or real-time processing is needed. Finally, reasoning about 3D meshes can only benefit from adapting to camera information when it is available [24,26].

In their pioneering work on HMR [22], Kanazawa *et al.* propose to predict SMPL mesh parameters and three parameters for weak-perspective reprojection given a cropped image containing a person. Different aspects of this approach have been improved since, including architectures [12, 26, 59], training techniques [25] and data enhancements [3, 21, 41]. The approach has also been extended to whole-body parametric models like SMPL-X [42], often with multiple crops centered on body, hands and face [7, 11, 37]. Multi-person inputs are typically handled with a two-step procedure: first running an off-the-shelf human detector, then applying a mesh recovery model on crops around each detected person. Conversely, ROMP [51] and PSVT [43] recover multiple human meshes in a single step using one-shot detectors. BEV [52] additionally predicts the relative depths of meshes. Accounting for intrinsic camera parameters has been shown to improve reprojection [24,26], especially when these differ between training and inference. Despite these advancements, no previous method has successfully integrated in a single model all four essential features: efficient multi-person processing, whole-body mesh recovery, location estimation

Table 1: Main features of Multi-HMR *vs.* the state of the art: Single-person methods rely on human detectors to process image crops around each person independently. Multi-person approaches detect humans and regress their properties using the same network. *Single-shot* refers to methods regressing the expected output without extracting or resampling features from different regions.

	Method	Whole Body	Single Shot	Camera Space	Camera Aware
Single-person	HMR [22]	✗	✓	✗	✗
	HMR2.0 [12]	✗	✓	✗	✗
	SPEC [24]	✗	✓	✗	✓
	CLIFF [26]	✗	✓	✗	✓
	PIXIE [11]	✓	✗	✗	✗
	Hand4Whole [37]	✓	✗	✗	✗
	PyMAF-X [58]	✓	✗	✗	✗
	OSX [27]	✓	✗	✗	✗
	SMPLer-X [4]	✓	✗	✗	✗
Det. + Single	3DCrowdNet [6]	✗	✗	✗	✗
Multi-person	ROMP [51]	✗	✓	✗	✗
	BEV [52]	✗	✓	✓	✗
	PSVT [43]	✗	✓	✓	✗
	Multi-HMR	✓	✓	✓	Optional

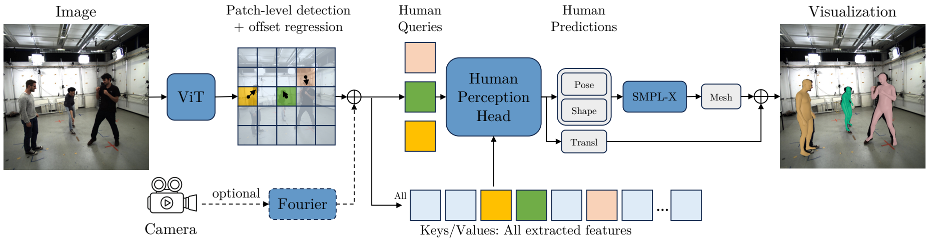


Fig. 2: Overview of Multi-HMR. A ViT backbone extracts image embeddings. Detection is conducted at the patch level with additional 2D offset regression. Each detected token serves as a query for a cross-attention-based head, called the Human Perception Head (HPH), which predicts pose and shape parameters, along with location in 3D space. Optionally, known camera parameters are embedded and added to each patch, represented as a Fourier encoding of the ray originating from the camera center.

in camera space and, optionally, camera-aware predictions. Please refer to Table 1 for a comparison to existing work.

In this paper, we introduce Multi-HMR, an efficient single-shot method that detects each person in a scene and regresses their pose, shape, and 3D location in camera space, using a whole-body parametric mesh model. Please see Figure 1 (left) for an example of prediction. Optionally, Multi-HMR can be conditioned on camera intrinsics if available. Figure 2 presents an overview of the model architecture. We use a standard Vision Transformer (ViT) [9] backbone to extract features from the input data, which allows us to benefit from recent advancements in large-scale self-supervised pre-training [5, 14, 40]. This differs from architectures like HR-Net [54] which are less common in the pre-training literature. We regress a person-center heatmap from the feature tensor produced by the backbone: for each input token, the model first outputs a probability that a person is centered on a point present in the corresponding input patch, as well as location offsets [63]. We introduce a prediction head called the Human Perception Head (HPH) that employs cross-attention. In this mechanism, queries correspond to the detected center tokens, while keys and values are drawn from all image tokens. It efficiently predicts pose and shape parameters of an expressive human model, namely SMPL-X [7], for a variable number of detections, while also regressing depths to place individuals within the scene. To improve 3D prediction by incorporating camera intrinsics, our model can optionally take camera parameters as input. These parameters are used to augment each token feature with Fourier-encoding of the corresponding camera ray directions before passing them to the prediction head.

Multi-HMR is conceptually simple: unlike most existing whole-body approaches, it does not rely on multiple high-resolution crops of the body parts for expressive models [7, 11, 37], or hand-designed components to place people in the scene [6, 51]. However, naively regressing SMPL-X parameters from a single token feature tends to under-perform on small body parts like hands. We find that incorporating expressive human subjects positioned close to the camera in the training data results in good performance across all body parts. We thus introduce the CUFFS (Close-Up Frames of Full-body Subjects) dataset, containing synthetic renderings of people with clearly visible hands in diverse poses.

We train a family of models with various backbone sizes and input resolutions. We evaluate performance on both body-only (3DPW [31], MuPoTs [33], CMU-Panoptic [20], AGORA-SMPL [41]) and whole-body expressive mesh recovery benchmarks (EHF [42], AGORA-SMPLX [41] and UBody [27]), see Figure 1 (right). The single-shot nature of the model allows for efficient inference. For instance, with a ViT-S backbone and 448×448 inputs, Multi-HMR is competitive on both body-only and whole-body datasets while being real-time, achieving 30 frames per second (fps) on a NVIDIA V100 GPU. Larger backbones and higher resolutions – up to a ViT-L backbone and 896×896 inputs – outperform the state of the art at the cost of slower but still reasonable inference speed (5 fps).

2 Related work

Multi-HMR primarily builds upon whole-body HMR and multi-person HMR. It also relies on synthetic datasets. We now review these three literatures.

Whole-body Human Mesh Recovery. There has been a recent surge of interest for whole-body mesh recovery from a single image [11,27,37,42], fostered in part by seminal work on improving whole-body parametric models. In particular SMPL-X [42] outputs an expressive mesh for the whole body given a small set of pose and shape parameters. The first approaches were based on optimization, *e.g.* SMPLify-X [42], but they remain slow and sensitive to local minima. Numerous learning-based methods were also introduced, but only in single-person settings [4, 7, 11, 37, 48, 58, 64]. This setting already poses significant challenges: hands and faces are typically low resolution in natural images, and capturing their poses hinges on subtle details. To overcome this, most approaches leverage a multi-crop pipeline: areas of interest – such as the face and hands – are cropped, resized and used to estimate the associated meshes which are aggregated into a whole-body prediction. In particular, ExPose [7] selects high-resolution crops using a body-driven attention mechanism; PIXIE [11] fuses body parts in an adaptive manner, and Hand4Whole [37] uses both body and hand joint features for 3D wrist rotation estimation. In contrast to these methods, Multi-HMR is single-shot, without high-resolution crops. More recently, OSX [27] introduced the first single-crop method for single-person whole-body mesh recovery. They leverage a ViT encoder, followed by a high-resolution feature pyramid, and use keypoint (*e.g.* wrists) estimates to resample features in their decoder head. SMPLer-X [4] employed a similar approach, training on numerous datasets. We depart from existing methods by i) tackling *multi-person* whole-body mesh recovery and ii) using a single-shot approach, with a non-hierarchical feature extractor.

Multi-Person Human Mesh Recovery. Most existing multi-person HMR methods [6, 12, 25, 44, 59] build upon a multi-stage framework: an off-the-shell human detector [15, 29, 47] is used, followed by a single-person mesh estimation model [23, 30, 57, 61] to process each detected human. This has two drawbacks: i) it is inefficient at inference time compared to a single-shot approach and ii) the pipeline cannot be learned end-to-end. This impacts final performance, in particular in cases of truncation by the image frame or person-person occlusions, a common scenario in multi-person settings. Following the seminal work of ROMP [51] which estimates 2D maps for 2D human detections, positions and mesh parameters, single-stage models have been proposed [43, 51, 52]. In particular, BEV [52] introduces an additional Bird-Eye-View representation of the scene to predict relative depth between detected persons and PSVT [43] improves performance using a transformer decoder. We follow the same single-shot philosophy as [43,51,52] but go beyond their settings by: i) tackling whole-body mesh recovery, ii) regressing the 3D location of each person in the camera coordinate system, and iii) incorporating camera intrinsics as an optional input. We also introduce an efficient cross-attention-based head, making Multi-HMR faster to train, efficient at inference and improving performance.

Synthetic data. Acquiring high-quality real-world ground-truth data at scale for human mesh recovery is costly, in particular when considering faces and hands expressions. A body of work [13, 53, 56] has explored the generation of large-scale synthetic data for human-related tasks. In this work, we experiment with BEDLAM [3] and AGORA [41], and confirm empirically that using large-scale synthetic data is beneficial for whole-body human mesh regression, compared to real-world data with pseudo ground-truth fits. We also propose a new synthetic dataset, CUFFS, which stands for Close-Up Frames of Full-body Subjects, designed to improve performance particularly on hands for one-stage whole-body prediction. It departs from existing ones in that it contains humans with diverse and clearly visible hand poses, seen from a limited distance, to allow fine details to be captured. Our experiments show that this type of training data is key to allow regressing whole-body meshes in a single shot.

3 Multi-HMR

We now describe our single-shot multi-person whole-body human mesh recovery approach. Given an input RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with resolution $H \times W$, our model, denoted $\mathcal{H}$, directly outputs a set of N centered whole-body 3D humans meshes $\mathbf{M} \in \mathbb{R}^{V \times 3}$ composed of V vertices, along with their corresponding root 3D locations $\mathbf{t} \in \mathbb{R}^3$ in the camera coordinate system:

$$\left\{\mathbf{M}_n + \mathbf{t}_n\right\}_{n \in \{1,\ldots,N\}} = \mathcal{H}(\mathbf{I}). \tag{1}$$

As preliminaries, Section 3.1 presents the 3D whole-body parametric model and the camera model that we use. We then detail the model architecture in Section 3.2 and the training losses in Section 3.3.

3.1 Preliminaries

Human whole-body mesh representation. We build upon the SMPL-X parametric 3D body model [7]. Given input parameters for the pose $\boldsymbol{\theta} \in \mathbb{R}^{53 \times 3}$

(global orientation, body, hands and jaw poses) in axis-angle representation, shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and facial expression $\boldsymbol{\alpha} \in \mathbb{R}^{10}$, it outputs an expressive human-centered 3D mesh $\mathbf{M} = \text{SMPL-X}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \in \mathbb{R}^{V \times 3}$, with $V = 10,475$ vertices. The mesh $\mathbf{M}$ is centered around a *primary* keypoint – in this work we choose the head as primary keypoint. It is placed in the 3D scene by putting the primary keypoint at the 3D location $\mathbf{t} = (t_x, t_y, t_z)$. For simplicity, let $\mathbf{x} = [\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}]$: the problem reduces to predicting $\mathbf{x}$ and $\mathbf{t}$ for all detected humans.

Pinhole camera model. We assume a simple pinhole camera model to project 3D points on the image plane. Ignoring distortion, it is defined by an intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times 3}$ of focal length f and principal point (p_u, p_v). We set the camera pose to the origin. We have:

$$\mathbf{K} = \begin{bmatrix} f & 0 & p_u \\ 0 & f & p_v \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \begin{cases} [c_u, c_v, 1]^T = (1/t_z) \cdot \mathbf{K}\ [t_x, t_y, t_z]^T \\ [t_x, t_y, t_z]^T = t_z \cdot \mathbf{K}^{-1}\ [c_u, c_v, 1]^T \end{cases}, \tag{2}$$

with $\mathbf{c} = (c_u, c_v)$ the 2D image coordinates of the projection of a 3D point $\mathbf{t}$ into the image plane. $\mathbf{K}$ can thus be used to backproject a 2D point into 3D given its depth t_z. We denote by $\pi_{\mathbf{K}}$ the camera projection operator and $\pi_{\mathbf{K}}^{-1}$ its inverse.

3.2 Single-shot architecture

Our method is summarized in Figure 2. We first encode images into token embeddings using a ViT backbone. These embeddings are used to detect humans and can optionally be combined with camera embeddings. Our proposed Human Perception Head is then employed to regress whole-body human meshes and depth for a variable number of detected humans.

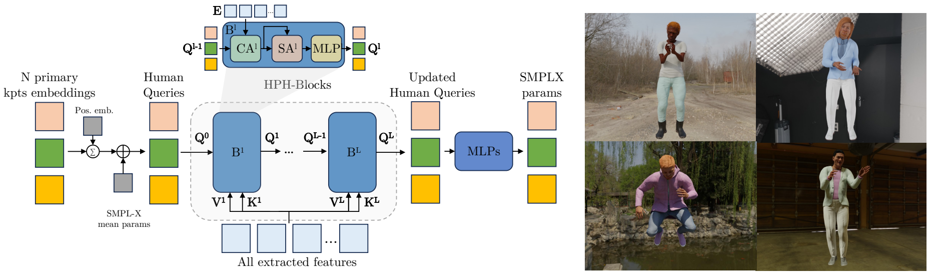
ViT backbone. The input RGB image $\mathbf{I}$ is encoded with a ViT backbone [9]. It is sub-divided into image patches of size $P \times P$, each embedded into tokens with a linear transformation and positional encoding. The set of tokens is processed with self-attention blocks into $\mathbf{E} \in \mathbb{R}^{H/P \times W/P \times D}$ with D the feature dimension. The ViT model keeps a constant resolution throughout so that each output token spatially corresponds to a patch in the input image.

Patch-level detection. To detect humans in the input image, we define a *primary keypoint* on human bodies, here the 3D keypoint of the *head* as defined according to the SMPL-X body model. For each patch index $(i, j) \in \{1, \ldots, H/P\} \times \{1, \ldots, W/P\}$, we predict if the patch centered at $\mathbf{u}^{i,j} = (u^i, v^j)$ contains a primary keypoint [63], with a score $s^{i,j} \in [0, 1]$ computed from the associated token embedding $\mathbf{E}^{i,j} \in \mathbb{R}^D$ using a Multi-Layer-Peceptron (MLP). At inference, we apply a threshold τ to the scores to detect patches containing primary keypoints:

$$\{\mathbf{u}_n\}_n = \{\mathbf{u}^{i,j} | s^{i,j} \geq \tau\}. \tag{3}$$

At train time, the ground-truth detections are used for the rest of the model.

Image coordinates regression. Detecting people at the patch level yields a rough estimation of the 2D location of the primary keypoint, up to the size of the predefined patch size P. We refine the 2D location of the primary keypoint



(a) Human Perception Head

(b) Samples from CUFFS

Fig. 3: (a) The token embeddings corresponding to the N detected primary keypoints are used as queries in a series of cross-attention blocks where keys and values correspond to the context provided by all image tokens. MLPs then predict the SMPL-X parameters (pose and shape) as well as the depth for each query. **(b)** Samples from our CUFFS synthetic dataset.

by regressing a residual offset $\delta = (\delta_u, \delta_v)$ from the center of a patch (u^i, v^j), using an MLP taking the corresponding token embedding $\mathbf{E}^{i,j}$ as input. The 2D coordinates predicted for the primary keypoint detected at patch location (i, j) are thus given by:

$$\mathbf{c}^{i,j} = [u^i + \delta_u, v^j + \delta_v]. \tag{4}$$

Human Perception Head (HPH). We predict human-centered meshes and depths estimations for all people detected in the scene in a structured manner and in parallel, by processing $\mathbf{E}$ with our Human Perception Head, built from cross-attention blocks [17], see Figure 3a for an overview. This design choice allows features corresponding to a person detection to attend information from all image patches before making a full pose, shape and depth prediction for this person. For a human detection n at patch location (i, j), we initialize a cross-attention query $\mathbf{q}_n = (\mathbf{E}^{i,j} \oplus \overline{\mathbf{x}}) + \mathbf{p}^{i,j}$, where $\mathbf{p}^{i,j}$ is a learned query initialization dependent on patch location, $\overline{\mathbf{x}}$ denotes the mean body model parameters, of dimension D' as in previous works [12, 25], and $\oplus$ denotes concatenation along the channel axis. Given N detections, the queries $\{\mathbf{q}_n\}_n$ are stacked into $\mathbf{Q}^0 \in \mathbb{R}^{(D+D')\times N}$ for efficient processing in parallel. The full feature tensor $\mathbf{E}$ is used as cross-attention keys and values. The queries are then updated with a stack of L blocks $\mathbf{B}^l$ (L=2 in practice), alternating between cross-attention layers ($\mathbf{CA}$) over queries and image features, self-attention layers ($\mathbf{SA}$) over queries, and an MLP:

$$\mathbf{Q}^l = \mathbf{B}^l \left[\mathbf{Q}^{l-1}, \mathbf{E}\right] = \mathbf{MLP}^l \left(\mathbf{SA}^l \left(\mathbf{CA}^l \left[\mathbf{Q}^{l-1}, \mathbf{E}\right]\right)\right). \tag{5}$$

The final outputs of the cross-attention-based module are given by $\mathbf{Q}^L \in \mathbb{R}^{(D+D')\times N}$ and viewed as a set of N output features, used to regress N human-centered whole-body parameters $\left\{\mathbf{x}_n\right\}_n$ with a shared MLP.

Depth parametrization. Following the monocular depth literature [34, 55], we predict the depth d in log-space, also called *nearness* [46] denoted η. We assume a *standard* focal length $\hat{f}$ and regress a *normalized* $\hat{\eta}$ from $\mathbf{Q}^L$ with an MLP:

$$\eta = \frac{\hat{f}}{f}\hat{\eta}, \quad d = \exp(-\eta). \tag{6}$$

This follows [10] which shows that this parametrization improves robustness to focal length changes. The depth d is used to back-project the 2D camera coordinates $\mathbf{c}$ using the camera inverse projection operator $\pi_{\mathbf{K}}^{-1}$ following Equation 2 to obtain the 3D location $\mathbf{t}$ of the primary keypoint.

Note that we directly supervise the *absolute* depth while most previous works [52] supervise the *relative* depth. This is made possible by the utilization of large-scale synthetic data, where absolute depth is known, as opposed to real-world data where only relative depth can be annotated. Our experimental results show the effectiveness of this simple strategy.

Optional camera embedding. If available, camera intrinsics $\mathbf{K}$ can be used as additional input to our model $\mathcal{H}$ which becomes $\mathcal{H}(\mathbf{I}, \mathbf{K})$. In more details, camera information may be integrated into the Human Perception Head at training and/or inference time. This is a desirable feature, but making it optional allows for i) processing images when it is not available, and ii) fairly comparing to the state-of-the-art methods that do no use this information.

We embed camera information by computing the ray direction [35] $\mathbf{r}_{i,j} = \mathbf{K}^{-1}[u_i, v_j, 1]^T$ from each patch center (u_i, v_j). The first two coordinates of the $\mathbf{r}_{i,j}$ vector are kept, and embedded into a high-dimensional space using Fourier encoding [35] to obtain a patch-level embedding $\mathbf{E_K} \in \mathbb{R}^{H/P \times W/P \times 2(F+1)}$, where F denotes the number of frequency bands. We concatenate features extracted using the vision backbone with camera embeddings to get $\mathbf{E} \coloneqq \mathbf{E} \oplus \mathbf{E_K}$.

3.3 Training Multi-HMR

Multi-HMR is fully-differentiable and trained end-to-end by back-propagation. We now discuss training losses. The symbol $\sim$ denotes ground-truth targets.

Detection loss. We project the ground-truth primary keypoint of each human present in the image using the camera projection operator $\pi_{\mathbf{K}}$, and construct a score map $\tilde{\mathbf{S}}$ of dimension $(W/P) \times (H/P)$ with $\tilde{s}^{i,j}$ equal to 1 if a primary keypoint is projected to the corresponding patch and 0 otherwise. Predictions are trained by minimizing a binary cross-entropy loss:

$$\mathcal{L}_{\texttt{det}} = -\sum_{i,j} \tilde{s}^{i,j} \log(s^{i,j}) + (1 - \tilde{s}^{i,j}) \log(1 - s^{i,j}). \tag{7}$$

Regression losses. All other quantities predicted by the model are trained with L_1 regression losses. We concatenate the offset from the patch centers $\tilde{\mathbf{c}}$, the body model parameters (pose, shape, expression) $\tilde{\mathbf{x}}$, following [12, 25], and the depth $\tilde{d}$ and minimize $\mathcal{L}_{\texttt{params}} = \sum_n \Big| \big[\mathbf{c}, \mathbf{x}, d\big] - \big[\tilde{\mathbf{c}}, \tilde{\mathbf{x}}, \tilde{d}\big] \Big|$. We also found it beneficial to minimize an L_1 loss for human-centered output meshes $\mathcal{L}_{\texttt{mesh}} =$

$\sum_n \left|\mathbf{M}_n - \tilde{\mathbf{M}}_n\right|$, as well as for the reprojection of the mesh onto the image plane $\mathcal{L}_{\texttt{reproj}} = \sum_n \left|\pi_{\mathbf{K}}(\mathbf{M}_n + \mathbf{t}_n) - \pi_{\mathbf{K}}(\tilde{\mathbf{M}}_n + \tilde{\mathbf{t}}_n)\right|$. The final training loss is thus:

$$\mathcal{L} = \mathcal{L}_{\texttt{det}} + \mathcal{L}_{\texttt{params}} + \lambda(\mathcal{L}_{\texttt{mesh}} + \mathcal{L}_{\texttt{reproj}}). \tag{8}$$

Synthetic whole-body CUFFS dataset. We introduce CUFFS[1], the Close-Up Frames of Full-body Subjects dataset, designed to contain synthetic renderings of people with close-up views of full-bodies with clearly visible hands in diverse poses, see Figure 3b. Using Blender [1], we render synthetic human models close to the camera, in poses sampled from the BEDLAM [3], AGORA [41], and UBody [27] datasets, using additional hand poses from InterHand2.6M [39] for increased diversity. Please refer to the supplementary material for more details. We render a total of 60,000 images. Simply adding this data during training improves the quality of hand pose predictions, without degrading other metrics.

Implementation details. By default, we use squared input images of resolution 448×448, with the longest side resized to 448 and the smallest zero-padded to maintain aspect ratio. We use random horizontal flipping as data augmentation. We initialize the weights of the backbone with DINOv2 [40] and experiment with Small, Base and Large ViT models as encoder. Please refer to the supplementary material for the full list of hyper-parameters and more implementation details.

4 Experiments

We first ablate training data and model architecture (Section 4.1), and then compare to the state of the art on body-only and whole-body HMR (Section 4.2).

Evaluation metrics. We evaluate the accuracy of the entire 3D mesh predictions with the per-vertex error (PVE), following [27,51,52], and also report it for specific body parts (hands and face). When the entire ground-truth mesh is not available, we report the Mean Per Joint Position Error (MPJPE) and the Percentage of Correct Keypoints (PCK) using a threshold of 15cm. We also report these metrics after Procrustes-Alignment (PA), and F1-Scores to evaluate detection. To evaluate the placement in the scene, we report the Mean Root Position Error (MRPE) [52] and the Percentage of Correct Ordinal Depth (PCOD) [60] metrics. For computational costs, we report inference time on a NVIDIA V100 GPU and the number of Multiply-Add Cumulation (MACs) using the *fvcore* library[2]. More details about the metrics are given in the supplementary material.

Evaluation benchmarks. For body-only benchmarks, we predict SMPL meshes from SMPL-X meshes using the regressor from [3], and follow prior work [27,37,43,51,52] in evaluating on 3DPW [31], MuPoTs [33], CMU [20] and AGORA [41]. For whole-body evaluation, we compare performance with prior work [11,27,37] on EHF [42], AGORA [41] and UBody [27]. We refer to the supplementary material for more details on datasets.

[1] `https://download.europe.naverlabs.com/ComputerVision/MultiHMR/CUFFS`

[2] `https://github.com/facebookresearch/fvcore`

Table 2: Architecture and training data are ablated on MuPoTs (PCK3D-All), 3DPW (MPJPE), EHF (PVE-All), EHF-H (PVE-Hands) and CMU (MPJPE). Default settings in grey . **(a)** We compare a ViT backbone to HRNet as well as our HPH with respect to a standard iterative regressor [22] ('Reg.'). **(b)** Training data type; 'Real'=MS-CoCo+MPII+Human3.6M,'A'=AGORA, 'B'=BEDLAM, and 'C'=CUFFS. When trained on 'C' only, we evaluate on single-person test sets only.

(a) Architecture

Backbone	Head	MuPoTs↑	3DPW↓	EHF↓	CMU↓
HRNet	Reg.	65.8	83.2	143.1	130.1
ViT-S	Reg.	70.1	80.2	90.6	118.1
HRNet	HPH	69.8	80.2	115.2	116.6
ViT-S	HPH	70.9	80.1	80.1	109.1
ViT-B	HPH	**76.3**	**73.5**	**55.3**	**97.2**

(b) Data

Data	MuPoTS↑	3DPW↓	EHF↓	EHF-H↓	CMU↓
Real	68.5	83.8	70.2	51.2	101.6
A+B	**76.3**	73.5	55.3	47.4	97.2
C	-	-	53.5	44.5	-
A+B+C	76.0	**72.9**	**49.8**	**40.5**	**96.5**
+Real	69.8	77.6	61.1	48.4	98.5

4.1 Ablations on model design and training data

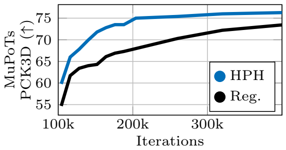
Default configuration. For the ablations, we use a ViT-B backbone with a HPH head composed of 2 blocks. We train only using synthetic the BEDLAM and AGORA datasets (but not CUFFS), without using the intrinsics as input. In each table the row of the default ablation configuration has a grey background.

Model architecture. We investigate several architectures in Table 2a. As most state-of-the-art single-shot methods (ROMP [51], BEV [52], PSVT [43]) use a HRNet [54] convolutional backbone, we evaluate both HRNet and ViT-S (as they have approximately equivalent parameter counts, 28.6M for HRNet and 21M for ViT-S) with either a vanilla iterative regression head [22] ('Reg.') or our proposed HPH. In both cases, the ViT-S backbone is beneficial and significant gains also come from our proposed HPH head, which validates our architecture. Scaling up the backbone (last row) further improves performance.

Training data. In Table 2b, we experiment with different types of training data. One source can be real-world datasets ('Real': MS-CoCo [28], MPII [2] and Human3.6M [16]), for which pseudo-ground-truth fits [36, 38] are obtained by minimizing the reprojection error of annotated 2D keypoints, but this remains inherently noisy. An alternative is to train on synthetic datasets such as AGORA [41] ('A') or BEDLAM [3] ('B') that have the advantage to be highly

Table 3: Ablation on the Human Perception Head (HPH). 'Reg.': parallel iterative regressors; HPH w/o **SA**: queries processed independently in HPH, *i.e.*, without self-attention, L: number of layers and H: number of heads. **(a)** Training convergence speed. **(b)** Impact of head choice. **(c)** Impact of HPH hyperparameters.

(a) Convergence



(b) Head architecture

Head	MuPoTS↑	3DPW↓	EHF↓
Reg. [22]	73.5	78.9	65.0
HPH w/o **SA**	74.5	76.4	63.2
HPH	**76.3**	**73.5**	**55.3**

(c) HPH Hyperparameters

L	H	MuPoTS↑	3DPW↓	EHF↓
2	8	76.3	73.5	55.3
2	4	76.5	74.0	54.8
4	8	78.5	72.4	51.3
8	8	**78.9**	**72.0**	**51.0**

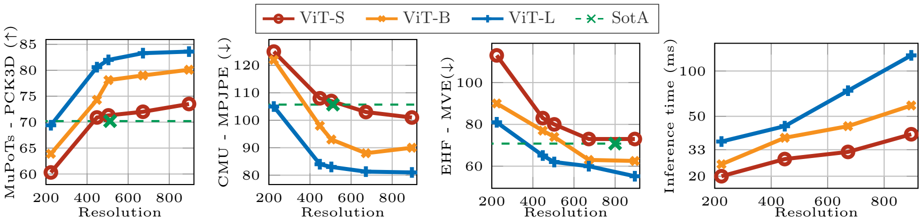


Fig. 4: Backbone-resolution-speed trade-off. We report the performance on MuPoTs, CMU and EHF using different backbone sizes and image resolutions. We also report the inference time (right).

scalable and to have perfect ground-truth. Recent work [3] has shown that state-of-the-art results can be achieved using synthetic training data only, despite an inherent sim-to-real gap. Our results confirm this finding as we obtain better results when training on large-scale synthetic data. When we add our synthetic CUFFS dataset ('C') we observe a significant boost in performance especially for metrics related to the hands (column EHF-H in the fourth row). However, when combining both real-world and synthetic datasets (last row), performance drops compared to training solely on synthetic data (penultimate row).

HPH architecture. In Table 3, we further compare different heads to regress the SMPL-X parameters. The baseline ('Reg.') uses a vanilla iterative regressor [25] applied to each detected feature token independently. 'HPH' converges faster (Table 3a) and performs better (Table 3b). 'HPH w/o **SA**' denotes a variant where queries are treated independently by removing **SA** blocks from the HPH, see Equation 5: treating queries together is beneficial (Table 3b). In Table 3c we experiment with different configurations of the HPH (number of layers 'L' and number of attention heads 'H'). Increasing the number of layers slightly improves performance but we favor the use of 2 layers for better efficiency.

Input resolution and backbone size. We evaluate the impact of the input image resolution for different backbone sizes (ViT-S, ViT-B, ViT-L) in Figure 4. Increasing the input resolution consistently brings performance gains across backbone sizes, at the cost of increased inference time (right). For body-only metrics, a ViT-L backbone at 448×448 inputs arguably offers a good performance *vs.* speed trade-off. Using higher resolutions may be more worthwhile for whole-body metrics; in particular, with a ViT-S or ViT-B backbone, high resolutions are critical to achieve competitive performance. This is to be expected as small details such as facial expressions and hand poses are easier to capture at high resolution – it motivated previous works [7,11,37] to extract specific high resolution crops for these parts. The largest backbone (ViT-L) at a 896×896 resolution takes approximately 120ms per image – without compressing or quantizing the network – which is fast compared to multi-stage methods (see Section 4.2).

Optional camera intrinsics. Integrating camera information is expected to improve accuracy when recovering and placing human 3D meshes in the scene. In Table 4a, we report results with different kinds of camera embeddings: computing

Table 4: Ablative study. Experiments on **(a)** the importance of the camera embedding type and **(b)** the sensitivity to the camera intrinsics in terms of human-centric reconstruction error and distance estimation error. $\hat{f}$: focal length normalization.

(a) Camera embeddings

	MuPoTS↑	3DPW↓	EHF↓
none	76.3	73.5	55.3
simple	74.8	75.3	56.8
rays	77.0	72.6	54.4
rays+$\hat{f}$	**78.8**	**71.3**	**53.1**

(b) Impact of optional intrinsics

FOV		Reconstruction↓			Distance (MRPE↓)		
Train	Test	MuPoTs	3DPW	CMU	MuPoTs	3DPW	CMU
60°	60°	76.3	73.5	97.2	1345	732	570
gt	60°	76.8	76.8	99.5	1512	731	595
gt	gt	**76.5**	**73.2**	**96.9**	**693**	**445**	**287**

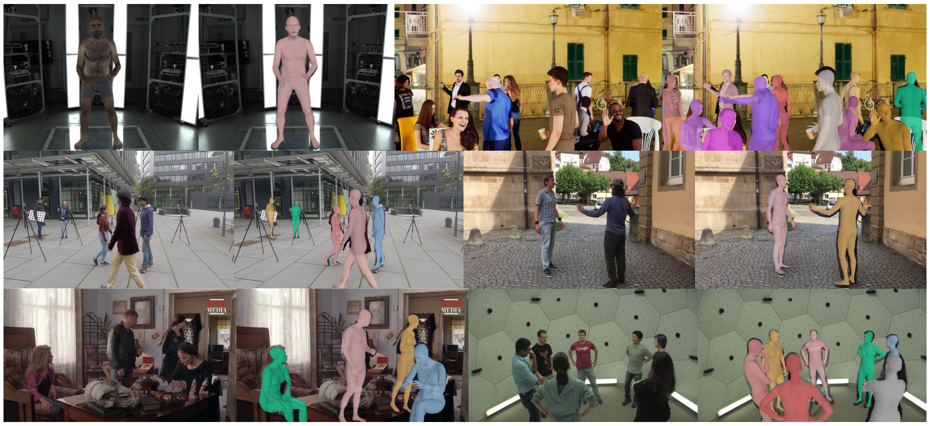


Fig. 5: Randomly sampled qualitative examples: input image and our results overlaid on it. Images from EHF and AGORA (top), MuPoTs and 3DPW (middle), UBody and CMU (bottom). See supplementary material for more visualizations.

simple embedding (where the flattened intrinsics matrix is fed to a linear layer) degrades performances compared to not adding camera embedding (*i.e.*, *none*) while adding *rays* brings a gain. When combined with focal length normalization $\hat{f}$, we observe a clear gain on all metrics. In Table 4b we report: performance with a fixed field of view (FOV) of 60°, like ROMP/BEV, for a model trained with intrinsics (row 1), and for a model trained without (row 2). Conditioning the model on camera intrinsics improves depth prediction accuracy (row 3), while reconstruction metrics which are centered on people are far less sensitive to this change. This validates the benefit of using intrinsics when available.

Other design choices. We present other ablations, *e.g.* on training losses and choice of primary keypoints, in the supplementary material.

Qualitative results. Figure 5 shows visualizations of some predictions.

4.2 Comparisons with the state of the art

No existing method is both multi-person and whole-body (Table 1). We thus compare either to multi-person approaches on body-only mesh recovery or to whole-body methods. In the latter case, our approach is single-shot, while others assume human detections, extract crops around each person, and process

Table 5: Comparison with state-of-the-art methods. As there is no other method that is both multi-person and whole-body, we compare separately to state-of-the-art approaches for **(a)** multi-person body-only mesh recovery, and **(b)** whole-body mesh recovery (all methods except Multi-HMR are single-person). For AGORA, we report performance for a single Multi-HMR setting due to restrictions of the evaluation system. † indicates a universal model which is not finetuned specifically for each benchmark.

(a) Body-only benchmarks

Method	**Res.**	**Single Shot**	**Backbone**	**3DPW**			**MuPoTs**		**CMU**		**AGORA**		
				PA-MPJPE↓	MPJPE↓	PVE↓	PCK-All↑	PCK-Matched↑	F1↑	MPJPE↓	F1↑	MPJPE↓	PVE↓
Body-only													
CRMH [18]	832	✓	RN50	-	-	-	69.1	72.2	0.92	143.2	-	-	-
3DCrowdNet [6]	Full		RN50	51.5	81.7	98.3	72.7	73.3	0.95	127.3	-	-	
ROMP [51]	512	✓	HR32	47.3	76.6	93.4	69.9	72.2	0.93	128.2	0.91	108.1	103.4
BEV [52]	512	✓	HR32	46.9	78.5	92.3	70.2	75.2	**0.97**	109.5	0.93	105.3	100.7
PSVT [43]	512	✓	HR32	45.7	75.5	84.9	-	-	**0.97**	105.7	0.93	97.7	94.1
Whole-Body													
Hand4Whole [37]	Full		RN50	54.4	86.6	-	-	-	-	-	0.93	89.8	84.8
OSX [27]	Full		ViT-L/16	60.6	86.2	-	-	-	-	-	-	-	-
SMPLer-X [4]	Full		ViT-L/16	51.5	76.8	-	-	-	-	-	-	-	-
SMPLer-X [4]	Full		ViT-H/16	48.0	71.7	-	-	-	-	-	-	-	-
Multi-HMR	896	✓	ViT-S/14	53.2	76.3	91.1	77.0	81.5	**0.97**	102.9	-	-	-
Multi-HMR	896	✓	ViT-B/14	46.7	70.9	86.9	79.4	84.6	**0.97**	94.6	-	-	-
Multi-HMR	896	✓	ViT-L/14	**41.7**	**61.4**	**75.9**	**85.0**	**89.3**	**0.97**	**77.3**	**0.95**	**65.3**	**61.1**
Multi-HMR	448	✓	ViT-L/14	43.8	64.6	79.7	77.8	84.1	0.96	84.0	-	-	-
Multi-HMR†	896	✓	ViT-L/14	46.9	69.5	88.8	80.6	86.4	**0.97**	97.5	-	-	-

(b) Whole-body benchmarks

Method	**Single shot**	**Backbone**	**EHF**						**AGORA**			**UBody-intra**					
			PVE↓			PA-PVE↓			PVE↓			PVE↓			PA-PVE↓		
			All	Hands	Face	All	Hands	Face	All	Hands	Face	All	Hands	Face	All	Hands	Face
Single person, per-body-part crops																	
ExPose [7]		HR32/RN18	77.1	51.6	35.0	54.5	12.8	5.8	217.3	73.1	51.1	-	-	-	-	-	-
FrankMocap [48]		RN50	107.6	42.8	-	57.5	12.6	-	-	55.2	-	-	-	-	-	-	-
PIXIE [11]		RN50	88.2	42.8	32.7	55.0	11.1	4.6	191.8	49.3	50.2	168.4	55.6	45.2	61.7	12.2	4.2
Hand4Whole [37]		RN50	76.8	39.8	26.1	50.3	10.8	5.8	135.5	47.2	41.6	104.1	45.7	27.0	44.8	8.9	2.8
PyMAF-X [58]		HR48	64.9	29.7	19.7	50.2	**10.2**	5.5	125.7	45.0	35.0	-	-	-	-	-	-
Single person, feature resampling																	
OSX [27]		ViT-L/16	70.8	53.7	26.4	48.7	15.9	6.0	122.8	45.7	36.2	81.9	41.5	21.2	42.2	8.6	2.0
SMPLer-X [4]		ViT-L/16	65.4	49.4	**17.4**	37.8	15.0	**5.1**	99.7	**39.3**	29.9	57.4	40.2	21.6	31.9	10.3	2.8
Multi-person, one forward pass																	
Multi-HMR	✓	ViT-S/14	50.0	43.3	24.4	36.8	14.4	5.8	-	-	-	56.9	35.7	18.9	23.8	9.9	2.5
Multi-HMR	✓	ViT-B/14	43.3	39.5	23.3	34.8	12.2	5.4	-	-	-	54.4	32.0	17.3	23.0	8.8	2.2
Multi-HMR	✓	ViT-L/14	**42.0**	**28.9**	18.0	**28.2**	10.8	5.3	**95.9**	40.7	**27.7**	**51.2**	**25.0**	**16.2**	**21.0**	**7.2**	**1.8**
Multi-HMR †	✓	ViT-L/14	**42.0**	**28.9**	18.0	**28.2**	10.8	5.3	-	-	-	54.0	27.5	17.0	22.8	8.0	2.4

each one independently. We report results with a 896×896 input resolution and without using camera intrinsics, with either a model finetuned for each benchmark as other methods do or a single universal model indicated by † (please refer to the supplementary material for additional information regarding finetuning).

Body Mesh Recovery. As most of these methods (ROMP [51], BEV [52] and PSVT [43]) use a 512×512 resolution, we also report results obtained at 448×448, which offers an excellent speed-performance trade-off. All these multi-person approaches are limited to body-only meshes. Multi-HMR outperforms existing work, with substantial gains across various metrics, even when using lower resolution input, smaller backbone or a universal model. At the same time, it also predicts hands poses and facial expressions (as evaluated next), which is not the case for other multi-person approaches.

Whole-Body Mesh Recovery. We evaluate our whole-body regression performance by comparing it against whole-body 3D pose methods [11,27,37]. All existing approaches are limited to the single-person scenario: they do not consider the

Table 6: Comparison to existing works for human depth estimation and inference cost. (a) Human depth estimation: we evaluate Multi-HMR without and with camera intrinsics information. **(b)** Comparison of inference cost for different number of humans N in an image between Multi-HMR (bottom) and the state of the art, which is limited to either multi-person but body-only methods (top), or single-person whole-body approaches thus requiring a human detector (middle).

(a) Depth estimation benchmark

	MRPE (↓)				PCOD (↑)	
Method	MuPoTs	3DPW	CMU	AGORA	MuPoTs	CMU
XNect [32]	639	-	-	-	-	-
ROMP [51]	1688	1060	679	-	91.2	97.1
BEV [52]	1884	1030	673	518	91.3	91.2
Multi-HMR						
w/o cam.	1125	522	355	421	95.1	98.5
w/ cam.	**514**	**318**	**110**	**396**	**97.9**	**99.5**

(b) Inference time and MACs

Method	SMPL-X	Params (M)	Time (ms) N=1	Time (ms) N=5	Time (ms) N=10	MACs (G) N=1	MACs (G) N=5	MACs (G) N=10
ROMP [51]		29.0	32.1	33.5	34.8	43.0	**43.6**	**44.2**
BEV [52]		35.8	36.6	37.8	39.1	48.6	48.9	49.9
Hand4Whole [37]	✓	77.9	73.3	366.5	733.0	**26.3**	98.3	188.3
OSX [27]	✓	102.9	54.6	273.5	546.0	94.8	440.8	873.5
Multi-HMR-S	✓	32.4	**28.0**	**28.6**	**28.8**	44.4	44.5	44.6
Multi-HMR-B	✓	99.0	38.0	38.9	39.0	143.9	144.2	144.4
Multi-HMR-L	✓	318.7	50.8	50.9	50.9	478.7	479.5	479.8

detection stage and the 3D positions in the scene, instead assuming predefined 2D bounding boxes around the person of interest. We report results in Table 5b. Multi-HMR is competitive with, or outperforms, previous whole-body methods, even when considering the universal model. In particular it obtains competitive performance on hands and faces (on par with or better than SMPLer-X [4], that is not single-shot). Overall, empirical results show that Multi-HMR predicts accurate hand and facial poses while also being multi-person.

Human depth estimation. In Table 6a, we compare the performance of our model in distance estimation, which uses simple depth regression, to the state of the art [32, 51, 52]. Prior works assume a fixed camera setting. For example, BEV [52] is competitive on AGORA-val but does not generalize as well to datasets with different camera parameters. The camera-aware variant of Multi-HMR provides accurate distance predictions across datasets and camera parameters, and the proposed approach still significantly outperforms the state of the art when camera intrinsics are not provided.

Inference cost. The number N of humans in an image defines the number of queries in the HPH head. With N=512, HPH takes 2.5ms *vs.* 2.3ms for N=5 on a NVIDIA V100 GPU. Other parts of the model are independent of N, thus our method scales well, as do other single-shot approaches (*e.g.* ROMP, BEV), see Table 6b. This is in contrast to multi-stage methods (*e.g.* Hand4Whole, OSX) which detect people,*e.g.* with YOLOv5 [19], and independently process their crops.

5 Conclusion

We presented Multi-HMR, the first single-shot method for multi-person whole-body human mesh recovery. It estimates accurate expressive 3D meshes (body, face and hands) and 3D positions in the scene, outperforming the state of the art for each sub-problem. Our model also adapts to camera information (*i.e.*, intrinsics) when available. Multi-HMR is conceptually simple: it relies on a vanilla ViT backbone and a newly introduced cross-attention-based head for predictions.

References

1. Blender. `https://www.blender.org/`
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
3. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR (2023)
4. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Wang, Y., Pang, H.E., Mei, H., Zhang, M., Zhang, L., et al.: Smpler-x: Scaling up expressive human pose and shape estimation. In: NeurIPS (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
6. Choi, H., Moon, G., Park, J., Lee, K.M.: Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In: CVPR (2022)
7. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV (2020)
8. De Santis, A., Siciliano, B., De Luca, A., Bicchi, A.: An atlas of physical human–robot interaction. Mechanism and Machine Theory (2008)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. Facil, J.M., Ummenhofer, B., Zhou, H., Montesano, L., Brox, T., Civera, J.: Camconvs: Camera-aware multi-scale convolutions for single-view depth. In: CVPR (2019)
11. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 3DV (2021)
12. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: ICCV (2023)
13. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: CVPR (2017)
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE trans. PAMI (2013)
17. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021)
18. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
19. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., TaoXie, Fang, J., imyhxy, Lorna, Yifu, Z., Wong, C., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, Sonck, V., tkianai, yxNONG, Skalski, P., Hogan, A., Nair, D., Strobel, M., Jain, M.: ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation (2022)
20. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015)

21. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In: 3DV (2020)
22. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
23. Kim, J., Gwon, M.G., Park, H., Kwon, H., Um, G.M., Kim, W.: Sampling is matter: Point-guided 3d human mesh reconstruction. In: CVPR (2023)
24. 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)
25. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
26. 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)
27. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR (2023)
28. 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)
29. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
30. Ma, X., Su, J., Wang, C., Zhu, W., Wang, Y.: 3d human mesh estimation from virtual markers. In: CVPR (2023)
31. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
32. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. ACM trans. Graph. (2020)
33. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV (2018)
34. Mertan, A., Duff, D.J., Unal, G.: Single image depth estimation: An overview. Digital Signal Processing (2022)
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
36. Moon, G., Choi, H., Chun, S., Lee, J., Yun, S.: Three recipes for better 3d pseudo-gts of 3d human mesh estimation in the wild. In: CVPR Workshop (2023)
37. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: CVPR Worskhop (2022)
38. Moon, G., Choi, H., Lee, K.M.: Neuralannot: Neural annotator for 3d human mesh training sets. In: CVPR Worskhop (2022)
39. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: ECCV (2020)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2023)
41. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: CVPR (2021)

42. 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)
43. Qiu, Z., Yang, Q., Wang, J., Feng, H., Han, J., Ding, E., Xu, C., Fu, D., Wang, J.: Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In: CVPR (2023)
44. Qiu, Z., Yang, Q., Wang, J., Fu, D.: Dynamic graph reasoning for multi-person 3d pose estimation. In: ACMMM (2022)
45. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., Malik, J.: On the benefits of 3d pose and tracking for human action recognition. In: CVPR (2023)
46. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3d appearance, location and pose. In: CVPR (2022)
47. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
48. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: ICCV (2021)
49. Salzmann, T., Chiang, H.T.L., Ryll, M., Sadigh, D., Parada, C., Bewley, A.: Robots that can see: Leveraging human pose for trajectory prediction. IEEE RAL (2023)
50. Shah, A., Mishra, S., Bansal, A., Chen, J.C., Chellappa, R., Shrivastava, A.: Pose and joint-aware action recognition. In: WACV (2022)
51. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021)
52. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting people in their place: Monocular regression of 3d people in depth. In: CVPR (2022)
53. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: CVPR (2017)
54. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE trans. PAMI (2020)
55. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: ICCV (2023)
56. Yang, Z., Cai, Z., Mei, H., Liu, S., Chen, Z., Xiao, W., Wei, Y., Qing, Z., Wei, C., Dai, B., Wu, W., Qian, C., Lin, D., Liu, Z., Yang, L.: Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In: ICCV (2023)
57. Yoshiyasu, Y.: Deformable mesh transformer for 3d human mesh recovery. In: CVPR (2023)
58. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. IEEE trans. PAMI (2023)
59. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021)
60. Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., Zhou, X.: Smap: Single-shot multi-person absolute 3d pose estimation. In: ECCV (2020)
61. Zheng, C., Liu, X., Qi, G.J., Chen, C.: Potter: Pooling attention transformer for efficient human mesh recovery. In: CVPR (2023)
62. Zhou, L., Meng, X., Liu, Z., Wu, M., Gao, Z., Wang, P.: Human pose-based estimation, tracking and action recognition with deep learning: A survey. arXiv preprint arXiv:2310.13039 (2023)

63. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: arXiv preprint arXiv:1904.07850 (2019)
64. Zhou, Y., Habermann, M., Habibie, I., Tewari, A., Theobalt, C., Xu, F.: Monocular real-time full body capture with inter-part correlations. In: CVPR (2021)