Multi-HMR 2: Multi-Person Camera-Centric Human Detection, Mesh Recovery and Tracking

Guénolé Fiche Philippe Weinzaepfel Romain Brégier Fabien Baradel

NAVER LABS Europe

https://github.com/naver/multi-hmr2

Fig. 1: Multi-HMR 2 detects humans and recovers their 3D meshes, placed in the scene, along with camera parameters. It also outputs per-human features that allow online tracking in videos, despite being trained only on still images.

Abstract. Most advances in human mesh recovery (HMR) have focused on pelvis-centered recovery, overlooking metric 3D localization and detection accuracy in the camera coordinate system – two key factors for real-world applications such as human–robot interaction and social scene understanding. Current evaluation protocols often ignore these aspects, emphasizing per-person, root-centered recovery rather than camera-space perception. As a result, existing approaches rely on fixed camera assumptions or handcrafted post-processing, limiting their robustness and practical deployment. We introduce **Multi-HMR 2**, a simple yet robust DETR-based framework for **Multi**-person camera-centric **H**uman detection, **M**esh **R**ecovery, and tracking. Multi-HMR 2 predicts a scene-consistent camera together with human meshes, enabling metric 3D localization without ground-truth intrinsics. Moreover, by distilling image-based memory features from SAM2, Multi-HMR 2 extends to tracking, achieving consistent identity association without video supervision. Despite its conceptual simplicity – no handcrafted components, no video input, and no ground-truth cameras – Multi-HMR 2 achieves state-of-the-art pelvis-centered performance while substantially improving detection accuracy and metric 3D localization.

1 Introduction

Perceiving humans from images is a core computer vision problem, with applications in diverse fields such as robotics, virtual reality, or automatic gesture analysis. Among human perception tasks, multi-person human mesh recovery (HMR) aims to detect, localize, and estimate the 3D mesh of every human in a scene. Although useful for understanding human interactions and for robotic navigation applications, this task is particularly challenging as one needs to deal with depth ambiguity, scene and inter-person occlusions, and computational cost constraints when recovering multiple humans simultaneously. Earlier works in multi-person HMR [9, 18, 44, 67] followed a two-step procedure consisting of: (1) detecting every human in the scene, typically with a pretrained object detection model [16, 33, 49, 61]; (2) estimating a mesh independently for every human by cropping the image around each person [6, 15, 21, 41, 68]. While enabling the use of powerful foundation models for object detection and single-person HMR, this paradigm presents several drawbacks: the inference cost grows with the number of individuals in the image, interactions between humans rely exclusively on global image features, and complex procedures are designed to localize humans in the 3D scene. To address these limitations, recent approaches propose to jointly recover all human meshes given an image using one-stage detectors.

Following the common practices in object detection, one-stage approaches to multi-person HMR can be divided into 2 categories: (1) In the spirit of CenterNet [71], some approaches rely on local image features to perform detection and feed the relevant ones directly into the decoder, (2) DETR-like models [7] have a fixed set of queries and need to perform query to target matching at training time and query selection at test time. Despite sometimes showing slower training convergence, DETR approaches typically better capture the global scene context as queries do not depend on local image features. Additionally, CenterNet-like approaches such as Multi-HMR [3] cannot handle images with close interactions when two humans are located in the same image patch, as the query would be the same. AiOS [54] and SAT-HMR [53] are DETR-based approaches. However, these methods make very strong assumptions about the camera model, with a fixed focal length of 5000 mm in [54] and a fixed field of view (FOV) of $\frac{\pi}{3}$ for [53]. This assumption leads to large errors in the estimate of 3D locations in the camera frame. Additionally, multi-person HMR methods [3, 53, 54] use parametric body models that can only represent adult bodies [34, 42]: this leads to an increased 3D localization error in the presence of children, which are predicted as adults at a long distance from the camera to compensate for the scaling error.

In this work, we introduce Multi-HMR 2, a multi-person HMR approach leveraging the Anny [5] parametric body model. Anny is open source and covers the human diversity — across ages (from infants to elders), body types, and proportions — with a unified model featuring interpretable body shape parameters. For detecting humans we rely on a DETR framework (see Figure 2): this makes Multi-HMR 2 a strong human detector, robust to heavy occlusions and to humans in close interaction. One of the limitations of DETR-based models is that training can be computationally expensive, depending on the modalities

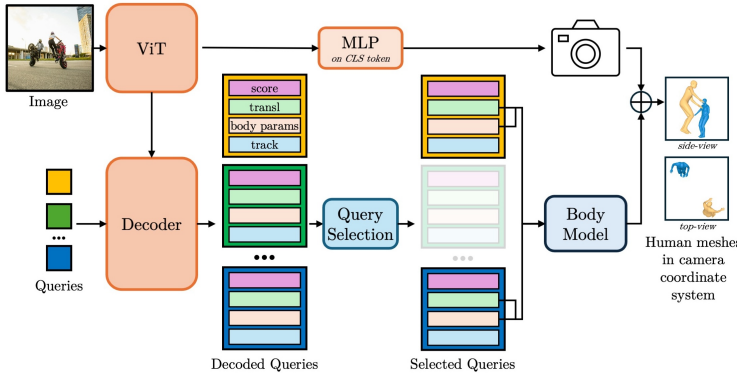

Fig. 2: Overview of Multi-HMR 2. Multi-HMR 2 follows a DETR-based architecture composed of an encoder that extracts image tokens and predicts camera parameters, and a decoder that cross-attends learned queries with image tokens before outputting score, 3D location, body parameters and tracking features for each query. A subset of queries is selected with an Hungarian matching at training or by thresholding the score at inference, and is fed to a body parametric model to obtain the final meshes.

used to associate the predictions with the targets in the Hungarian matching algorithm. In contrast with [53, 54], we propose to match predictions with targets based solely on the 2D location of humans, avoiding costly forward and backward passes through the parametric body model for all queries. We show that this does not affect the HMR performance while allowing us to use additional sources of training data without ground-truth meshes and significantly reducing the compute needs for training. Multi-HMR 2 is trained in around a week on a single GPU.

By contrast to [53,54], we additionally propose to estimate the camera intrinsic parameters. This allows us to localize humans much more accurately in the camera space, both in terms of absolute distance to the camera and relative distance between humans. Combined with the ability to model children, Multi-HMR 2 presents unprecedented capability in metric-scale recovery of human-centered scenes from a single image.

In addition to 3D properties, we propose to regress features distilled from the SAM2 [48] memory encoder for each human, that can be used to track humans by associating detections across frames. This makes Multi-HMR 2 the first approach for human tracking that can be trained without any video or mesh sequence.

We evaluate Multi-HMR 2 on in-the-wild HMR benchmarks [22, 37], humans in close interaction [23,66], reprojected keypoints [29], and on the tracking task [10]. We show that despite its simplicity, Multi-HMR 2 achieves results superior or comparable to the SOTA multi-person HMR methods.

In summary, we make the following contributions:

- We introduce Multi-HMR 2, a multi-person HMR method based on the inclusive Anny body model.

- Multi-HMR 2 is based on the DETR framework, making it robust to overlapping humans. Our novel matching strategy based on 2D location makes Multi-HMR 2 fast to train and trainable with images without 3D annotations.
- Multi-HMR 2 can use the ground-truth camera parameters or predict them, allowing highly-accurate human localization in the scene.
- We distill SAM2 features, allowing Multi-HMR 2 to track humans by associating detections across frames without any training on video data.

2 Related work

Multi-person HMR. Human mesh recovery, introduced in [21], is the task of regressing human meshes given an image. In this work we focus on parametric methods, that predict the parameters of a human body model [5,34,42,63]. While most methods focus on the single-person scenario and assume that their input is an image tightly cropped around the human of interest, more and more works shift to the multi-person HMR task. Multi-person HMR methods aim to detect and localize humans in the 3D scene in addition to regressing human meshes. Earlier approaches were multi-stage [9, 18, 44, 67] and relied on off-the-shelf human detectors [16, 33, 49, 61] and single-person HMR methods [6, 15, 21, 41, 68]. However, this multi-stage paradigm lead to an inference time linearly growing with the number of subjects in the image, and ignored the spatial relationships and interactions between humans. Following the seminal work of ROMP [55], some single-stage approaches were proposed [3, 43, 56]. These approaches are inspired by the CenterNet [71] paradigm: they first detect humans and then use their local image features as queries to predict meshes. Recently, PromptHMR [60] united the CenterNet and multi-stage paradigms by enriching Multi-HMR with ground-truth or off-the-shelf detections, segmentation, and textual information. However, this kind of approaches typically ignores the global context, as queries contain most of the necessary information to predict human meshes. Additionally, they are not optimal for handling overlapping humans. For these reasons, recent work explores DETR-like approaches for multi-person HMR [53, 54].

In this work, we propose a new DETR-based approach to multi-person HMR. The DETR framework allows end-to-end training, better takes into account the global image context, and provides an efficient and elegant way to deal with detections, especially in occlusion scenarios.

DETR for human perception. The DETR architecture [7] was originally proposed for object detection. It consists of feeding a set of learned queries to the decoder. Each query will result in a prediction, with a given confidence value. During training, predictions are matched with the ground truth using a Hungarian algorithm. At inference, queries are selected using the predicted confidence value. Many methods built upon DETR [8, 17, 32] improved its convergence, and first applications to human perception were made in 2D pose estimation [30,51,65]. Starting from ED-Pose [65], AiOS [54] was the first DETR-based

approach to multi-person HMR. AiOS first localizes subjects with a human token and introduces joint-related tokens to predict expressive human meshes [42]. However, although being one-stage, AiOS relies on crops to predict fine-grained details such as hands and facial expression, and it assumes a focal length of 5000 mm, making the global localization prediction inaccurate. SAT-HMR [53] extends Multi-HMR [3] to the DETR framework with scale-adaptive image features. SAT-HMR achieves performance comparable with [3,54] in pelvis-centered metrics with a minimal inference cost, however, its predictions in the camera coordinates system lack accuracy as it relies on cameras with a fixed FOV.

Multi-HMR 2 focuses on improving existing multi-person HMR methods in terms of 3D localization in the camera frame. For that, we propose an adaptive model that can leverage ground-truth camera parameters or predict the FOV. We show that Multi-HMR 2 achieves SOTA performance in pelvis-centered metrics while significantly improving scene-centered metrics.

Human tracking. Multi-person HMR can be extended to videos, which requires tracking humans across frames to leverage the temporal context. Most approaches to human tracking work in a tracking-by-detection setting, which means that they detect objects in every frame independently and then associate them. Earlier approaches were based on 2D locations and keypoint features [14, 62]. Incorporating HMR predictions into the tracking pipeline significantly improves accuracy [45, 46], and [15] show that stronger detections and human mesh recovery estimates lead to even better tracking. Recently, CoMotion [39] proposed a tracking-by-attention method that detects new objects and updates existing tracks jointly, and [31] proposed a DETR-based approach that considers the predictions of a given query across time as a motion prediction.

While prior works all rely on video or human motion data for training, we propose to distill SAM2 features, that can be used for tracking by associating detections across frames while being trained only on still images.

3 Method

Overview. Figure 2 shows an overview of the Multi-HMR 2's architecture, that given a single RGB image, detects humans in the scene and regress their meshes in the camera coordinate system. The model is based on the DETR [7] paradigm for object detection: the image is encoded with a ViT. A decoder based on cross-attention blocks with respect to the encoded tokens then processes a set of learned queries. These decoded queries are then fed to different MLP heads: a first one predicts a detection score, a second one predicts global location, a third one predicts the parameters of a parametric human body model, and a last one predicts a feature that can be used for tracking. We additionally regress the camera intrinsics from the *cls* token of the encoder.

DETR-like detectors have the advantage of elegantly dealing with overlapping humans, in contrast to CenterNet-like approaches [3,56] or two-stage methods [15]. While they might be slow to train, in particular if all queries are forwarded to the body model during training, we propose to perform the matching

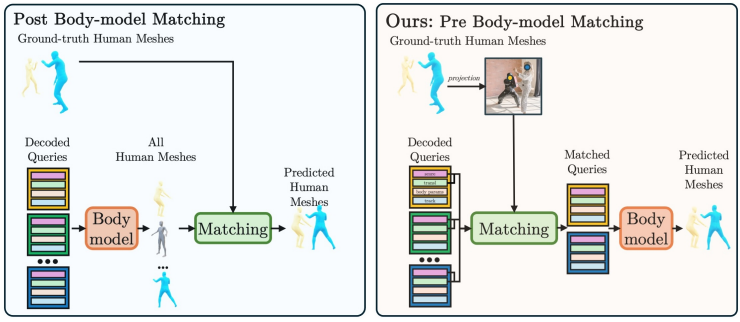

Fig. 3: Matching queries and ground-truth humans during training. Prior works [53, 54] fed all queries to the body model to perform matching using all joints (left). In contrast, we perform the matching using the location only (right): only the matched queries require a forward pass through the body model, significantly reducing computational cost and accelerating training.

considering the body location only, avoiding this problem (see Figure 3). By using only the 2D localization in pixels, we are also able to train using datasets annotated with 2D keypoints alone [29].

As parametric body model, we rely on the recent Anny [5], for its strong ability to model human diversity (in terms of ages covering babies to elders, genders, weight, *etc.*). As most other parametric models [42, 63], Anny encodes human meshes, including face and hands, using only a few bone 3D orientations, namely 163 and a set of 6 interpretable body shape parameters.

As feature relevant for tracking, we propose to distill the flattened feature maps of the SAM2 [48] image and memory encoder: this can be trained from still image supervision, *i.e.*, without any video annotated with human meshes. This stands in contrast to classical feature obtained from appearance cues [14, 46].

We now detail the components of our model.

Image Encoder. The input image is processed with a ViT-Large encoder [11]: the image is split into patches, which are fed to a series of transformer blocks [57]. A linear layer maps the obtained image features to the dimension of the decoder. While most HMR images process only square images, by padding the borders, we train our model with various aspect ratios. We crop the images during training to randomly chosen aspect ratio while the largest dimension is set to 768 pixels. The crops are centered to preserve the centering of the principal point. At inference, the image is simply resized such that the largest dimension is 768 and the smallest side a multiple of the patch size.

Camera intrinsics prediction. We use a Multi-Layer Perceptron (MLP) head to predict the camera intrinsic parameters from the *cls* token of the ViT encoder. More precisely, we regress the vertical FOV, assuming similar focal length along horizontal and vertical axis, and a principal point at the center of the image.

Decoder. The Multi-HMR 2's decoder follows a DETR-like architecture. A set of 100 learned human queries is decoded through 8 decoder blocks, each com-

posed of a multi-head self-attention layer between queries, a multi-head cross-attention layer with image tokens, and an MLP. The decoded queries will then be given as input to multiple prediction heads, each responsible for the prediction of a given modality, as detailed below.

Detection score. Our model predicts a detection score for each decoded query. This score will be used in the matching process during training, while at inference time, we only keep queries with a score above a given detection threshold.

Location prediction. Regressing the 3D location of humans in the camera space is particularly challenging, especially when focal length is not fixed. We decompose the global location estimation into 2 simpler problems. First, we estimate the 2D location of a human in pixel space. Following the monocular depth literature [12,38,47] and prior works in multi-person HMR [3,60] we then predict the nearness, that corresponds to the depth in log-space. We can then recover the 3D location by inverse projection using our predicted depth and 2D location, and estimated camera intrinsics.

Matching queries and ground-truth human meshes during training. Since our model is based on the DETR framework, predictions need to be matched with the ground-truth meshes during training. Unlike prior DETR-like HMR methods [53, 54], we propose to do the Hungarian matching based solely on the predicted confidence and 2D location, before predicting meshes. Let us denote $\hat{y}=\{\hat{l}_i, \hat{c}_i\}_{i=1}^{N}$ the N predicted locations in pixels and confidence values, and $y=\{l_i\}_{i=1}^{M}$ the M target locations. Let x_{ij} be the binary variables indicating if the prediction i matches the target j. We solve the following problem:

$$\underset{\{x_{ij}\}}{min} \sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}\left(\|\hat{l}_i - l_j\|_1 + FC(\hat{c}_i)\right) \tag{1}$$

$$\text{s.t.} \quad \sum_{j=1}^{M} x_{ij} \leq 1 \quad \forall i=1...N, \qquad \sum_{i=1}^{N} x_{ij} = 1 \quad \forall j=1...M,$$

with $FC(\hat{c}_i) = -\alpha(1-\hat{c}_i)^{\gamma}\log(\hat{c}_i)$ a focal loss [28] with α=0.25 and γ=2.

This 2D-based matching strategy offers two key advantages. First, it greatly reduces the computational cost, since only matched queries are passed through the differentiable parametric body model, while unmatched queries are discarded early as shown in Figure 3. A forward-backward pass takes 0.35s with our matching strategy vs. 0.48s with the standard one. It also uses much less memory (7GB vs. 27GB), enabling larger batch sizes and further speed gains. Second, it enables training with images annotated solely with 2D keypoints or bounding boxes, as no 3D mesh supervision is required for the matching step. In Table 6 of Section 4, we demonstrate through an ablation study that this strategy maintains comparable accuracy to 3D-based matching while significantly improving training efficiency and flexibility across heterogeneous supervision sources.

HMR prediction. For each matched query, we predict the Anny shape parameters and the joint angles in the form of 6D rotations [72] that correspond to relative angles along the kinematic tree. The Anny body model then outputs the mesh vertices and joint locations in a differentiable fashion.

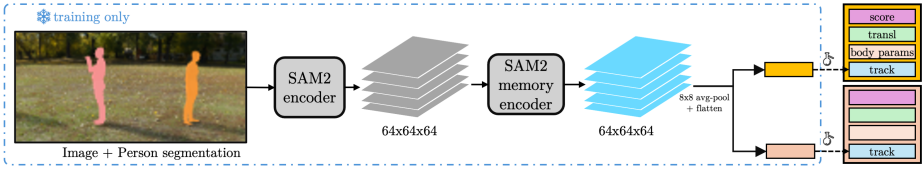

Fig. 4: Target tracking feature. We obtain a target feature for each person using SAM2. The image and ground-truth mask are fed into SAM2, and the average-pooled memory encoder output is flattened into a feature vector.

Regressing track features. We aim to also obtain an appearance feature per human that would allow to associate detections across frames in videos at inference, thus allowing to track people. One possible solution for that is to build an explicit texture map [15, 45, 46] but this might suffer from inaccurate mesh estimation or varying illuminations. An alternative is to rely on contrastive learning, inspired by the person re-identification literature [1,35], but this would require pairs of images; or distillation of a strong already trained model. However, these models are known to suffer from dataset biases [69, 70]. In this work, we propose to distill the knowledge from the SAM2 foundation model for image and video segmentation [48] (see Figure 4).

SAM2 processes each video frame independently, but it maintains a memory bank that stores short-term spatio-temporal information from recent frames. For a new frame, the model first encodes the image into tokens. These tokens then interact, via cross-attention, with the memory bank, which holds features from previous frames. A mask decoder uses the resulting features to generate the segmentation mask for the current frame. After producing the mask, SAM2 encodes both the mask and the memory-enhanced features from the current frame to create new feature representations. These features are added to the memory bank and correspond to a 64×64 spatial grid of features, each of dimension 64. The memory bank contains the set of features from the last 7 frames at maximum.

In our case, given a training image with ground-truth human segmentations, we build a tracking feature (as shown in Figure 4) by (a) encoding the image with the SAM2 image encoder thus obtaining an image feature map, (b) feeding that image feature map with the ground-truth segmentation to the memory encoder to obtain a memory feature map, (c) compressing this $64 \times 64 \times 64$ representation into a 4096-dimensional vector using an 8×8 average pooling and a flattening operation. This feature can be computed for every human in training images as long as ground-truth segmentations are available, either with annotations [29] or thanks to the synthetic nature of many datasets [4, 5].

Training losses. Training losses include a cross-entropy for the predicted confidence, and L_1 losses for the predicted 2D locations, depths, pose parameters, shape parameters, 3D vertices and joint positions centered on the pelvis, reprojected 2D joint coordinates using the ground-truth or predicted camera, and

regressed track features. Following [41], we use an asymmetric loss for the predicted FOV, which penalizes overestimated values more.

In order to improve the stability of training, we additionally use a stable loss for 2D joints based on the reprojection error on a sphere. The inverse projection operator $\Pi_{\boldsymbol{K}}^{-1}$ that maps 2D camera coordinates to 3D directions on the half-sphere in front of the camera is continuous and differentiable, contrary to $\Pi_{\boldsymbol{K}}$. We therefore minimize the angular error between the predicted 3D directions and target rays:

$$\mathcal{L}_{\text{2D angle}} = \frac{1}{s_{\boldsymbol{K}}} \sum_i \angle(\boldsymbol{x}_i, \Pi_{\boldsymbol{K}}^{-1}(\hat{\boldsymbol{u}}_i)) \tag{2}$$

where $s_{\boldsymbol{K}}$ is a scaling factor defined as the camera diagonal field of view angle. It is stable except for points located at the camera center, where the angle is undefined; we thus ignore points within a 0.1m radius around the camera center when computing this loss.

Dealing with heterogeneous training data. Our training relies on a mixture of synthetic data, pseudo-ground-truth (pseudo-GT) meshes on real images, and real images annotated with only 2D keypoints. We adapt the supervision strategy for each data type based on its reliability and annotation quality. For *synthetic data*, which provides accurate ground-truth meshes and cameras, we supervise all prediction heads: camera parameters, detection confidence, 2D locations, depth, body pose, shape, vertices and their 2D reprojections. For *pseudo-GT real images*, we supervise only the 3D body pose and shape, but not the detection confidence, since the annotations are sparse and do not cover all people in the scene. This prevents penalizing valid detections that lack annotations. Finally, for *real images with 2D keypoint annotations only*, we supervise the detection heads and the 2D re-projection of certain joints (*i.e.* MS-COCO). This encourages alignment between projected 3D keypoints and annotated 2D joints, enabling training even without mesh supervision. This adaptive supervision strategy allows Multi-HMR 2 to effectively combine heterogeneous datasets with varying annotation quality, ensuring robustness and generalization across both synthetic and real data.

Training details. The encoder weights are initialized with DINOv3 [52], while other parts of the model are trained from scratch. We use the Adam optimizer [24] for training. As training data, we use the Anny-One [5] and BEDLAM [4] synthetic datasets, as well as real images from MSCOCO [29], MPII [2] and AIC [13] with the pseudo-ground truth meshes from CameraHMR [41]. For all datasets except Anny-One, that is already in the Anny format, we fit the Anny model to existing SMPL [34] annotations. This model pretrained exclusively with 3D annotations is denoted **Multi-HMR 2.a** in the experiments section.

One of the weaknesses of the pseudo-ground truth meshes is that they are sparse, which can lead to poor detection performance on challenging in-the-wild images. We experiment using additional 2D large-scale pseudo-ground truth on MSCOCO [29] and Openimages [26], obtained using a strong detector [61] and 2D pose estimation method [64]. However, we observe that naively mixing 2D

and 3D supervision alters the pose prior learned during 3D training. In particular, when lower-body joints are occluded, the model frequently predicts sitting poses. We observe a similar behavior in SAT-HMR [53], suggesting that this issue is not specific to the architecture but rather is a consequence from sparse 2D supervision. To mitigate this undesired effect, we propose to regularize the finetuning stage towards the predictions of the Multi-HMR 2.a model. Specifically, we add consistency L_2 losses on pose parameters and 3D joints, together with a reprojection loss on COCO joints when 3D annotations are available. The finetuned model is denoted as **Multi-HMR 2.b** in the experiments section.

For the target tracking features obtain from SAM2 [48], we use its Hiera-L version [50].

Inference on images. At inference, the image is padded to the closest aspect ratio used for training, and resized so that its largest dimension is 768. Instead of doing Hungarian matching on the 2D location, we keep the queries with a confidence value greater than 0.4. Optionally, a non-maximum-suppression is done based on the 3D location.

Tracking in videos. In the case of videos, we process the frames one by one and detect humans and their meshes with our model. Following methods based on per-frame appearance features, *e.g.* [46], we build a similarity matrix between human detections in the current frames and active human tracks from previous frames and finally use the Hungarian matching algorithm to associate an ID to each detection in the current frame. We now give more details about the similarity score we use.

We first build a similarity matrix between all new detected humans and all active human tracks based on the predicted features distilled from SAM2. To this end, we keep a bank of features seen in the past 50 frames, together with their assigned IDs. For each new feature, we use K-Nearest-Neighbors (KNN) search to retrieve the most similar (K=35) in the bank according to the L1 distance, apply a softmax function with a temperature parameter of t=15, and aggregate the score for each human ID.

Second, for each human ID, we keep track of the last 8 positions and orientations of the pelvis keypoint in 3D space, and we use a ridge regression, to predict the expected position and orientation of the pelvis in the current frame. Let $\hat{p}_j, \hat{n}_j, \hat{\theta}_j$ be the predicted pixel, nearness and orientation respectively for a given human ID j. Given a new human detection i and its pelvis pixel p_i, nearness n_i and orientation θ_i, we build a similarity matrix between each new detection and every human ID already assigned using the following formula: $e^{-\|p_i - \hat{p}_j\|/t_p} e^{-\|n_i - \hat{n}_j\|/t_n} e^{-\|\theta_i - \hat{\theta}_j\|/t_\theta}$ with t_p=1, t_n=3 and t_θ=20.

The two similarity matrices are combined to obtain a score per association using a weighted sum with weights of 1 (resp. 8) for the feature (resp. pelvis) similarity. Once the similarity matrix is computed, we use an Hungarian matching algorithm to make the association. We stop a track when it had no association in the last 50 frames. We start new tracks when some new detections have no association with a similarity below 0.69.

Table 1: Evaluation on HMR benchmarks (3DPW and EMDB).

Method	Camera	3DPW				EMDB			
		PA-MPJPE ↓	MPJPE ↓	PVE ↓	TA-PVE ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Abs-PVE ↓
AiOS [54]	Fixed	45.0	68.8	90.9	774.3	63.3	90.6	108.1	11962.1
SAT-HMR [53]	Fixed	52.7	81.0	94.5	174.5	71.0	112.9	126.7	1093.0
Multi-HMR 2.a	Predicted	**40.5**	68.2	80.1	170.1	**48.3**	**68.5**	**78.7**	**273.6**
Multi-HMR 2.b	Predicted	41.8	**65.2**	**78.1**	**161.9**	52.5	71.0	83.6	374.8
Multi-HMR [3]	GT	46.9	69.5	88.8	187.2	48.5	73.7	87.1	168.1
Multi-HMR 2.a	GT	**40.2**	67.7	79.9	149.8	**48.3**	**68.5**	**78.7**	**129.4**
Multi-HMR 2.b	GT	41.6	**64.5**	**77.7**	**145.6**	52.5	71.0	83.6	131.7

Table 2: Evaluation on interaction HMR benchmarks (Hi4D and Harmony4D).

Method	Camera	Hi4D				Harmony4D			
		MPJPE ↓	Pair-PA-MPJPE ↓	TA-PVE ↓	F1-score ↑	MPJPE ↓	Pair-PA-MPJPE ↓	TA-PVE ↓	Recall ↑
AiOS [54]	Fixed	72.9	239.0	694.4	98	**130.5**	304.8	3454.8	91
SAT-HMR [53]	Fixed	88.2	85.5	132.4	**100**	140.3	144.0	676.3	**97**
Multi-HMR 2.a	Predicted	**61.7**	**79.5**	**114.9**	**100**	131.8	**99.8**	**231.2**	**97**
Multi-HMR 2.b	Predicted	65.9	82.9	123.5	99	132.4	105.4	382.1	96
Multi-HMR [3]	GT	69.2	82.2	116.2	98	146.0	140.8	388.1	93
Multi-HMR 2.a	GT	**61.7**	**78.3**	**113.1**	**100**	**128.4**	**96.2**	**233.6**	**97**
Multi-HMR 2.b	GT	65.9	80.5	119.5	99	128.7	96.7	274.6	**97**

4 Experiments

We compare Multi-HMR 2 to the state of the art in Section 4.1 and then ablate its main components in Section 4.2.

Evaluation benchmarks. Following recent work in HMR, we evaluate the in-the-wild HMR performance of Multi-HMR 2 on the 3DPW [37] and EMDB [22] benchmarks. We also use the Hi4D [66], Harmony4D [23], and CMU-Panoptic [20] test sets, which focus on people in close interaction, to evaluate the detections under occlusion and the capacity of the model to capture the relative position of humans in the scene. We evaluate the reprojection on the MSCOCO [29] dataset, and for the tracking task, we test on the PoseTrack21 [10] validation set.

Metrics. We report standard HMR metrics in mm: the per-vertex error (PVE) measures the Euclidean distance between the vertices of the predicted mesh and the ground truth centered on the pelvis, the mean-per-joint-error (MPJPE) focuses on joints extracted from the mesh, and the Procrustes-aligned MPJPE (PA-MPJPE) computes the error on joints after a rigid transformation that minimizes the error for each mesh. In order to evaluate the location estimate in the camera coordinates system we also compute a Abs-PVE (Absolute PVE) that is not centered on the pelvis, and the TA-PVE (Translation-Aligned PVE) that aligns the location of the scene and can only be computed when at least 2 humans are detected. On Hi4D and Harmony4D, we also evaluate the interaction with the Pair-PA-MPJPE that applies a single Procrustes-alignement to both humans altogether instead of doing it for each mesh separately, and compute the F1-score and Recall to evaluate the detection under occlusions. On the CMU-Panoptic dataset we compute the mean-root-positional-error (MRPE) to

Table 3: Evaluation on CMU-Panoptic.

Method † *use official code+ckpt*	haggling	mafia	ultimatum	pizza	mean	
	N-MPJPE ↓				N-MPJPE ↓	N-MRPE ↓
3DCrowdNet [9]	115.4	143.1	136.6	142.7	134.0	-
BEV [56]	93.5	106.9	116.6	129.1	112.9	-
PSVT [43]	91.4	100.9	118.8	124.8	109.0	-
DPMesh [74]	97.2	109.8	114.3	120.5	110.4	-
Scene [36]	84.5	-	108.9	133.2	108.9	217.1
SAT-HMR† [53]	107.4	134.2	126.9	126.8	123.8	83.7
Multi-HMR† [3]	<u>73.1</u>	<u>85.8</u>	<u>94.0</u>	**84.6**	<u>84.4</u>	<u>58.1</u>
Multi-HMR 2.a	**68.9**	**81.3**	**82.7**	<u>88.9</u>	**81.5**	**26.9**

Table 4: Comparison of predicted FOV (in degrees).

Method	3DPW	EMDB
Perspective fields [19]	14.0	12.8
Ctrl-C [27]	<u>5.6</u>	5.4
WildCamera [73]	8.2	**2.3**
SPEC-camcalib [25]	8.8	5.9
HumanFoV [41]	**5.0**	5.7
Multi-HMR 2.a	6.4	<u>4.0</u>

capture the accuracy of the translation, and report metrics normalized by the F1-score to account for detection. For evaluation on HMR tasks, following prior works, we match the predictions with the targets using an Hungarian matching based on the 2D joint locations. Metrics on MSCOCO are the keypoint average precisions (AP) and recalls (AR). To evaluate tracking, we report the ID F1-score (IDF1) that evaluates how well the tracker maintains correct object identities over time, and the multiple object tracking accuracy (MOTA) that takes into account missed detections, false positives, and identity switches. We use the evaluation code from [39], that correctly accounts for the regions to ignore.

4.1 Comparison with the state of the art

We evaluate Multi-HMR 2 on in-the-wild HMR (Table 1), interaction recovery (Table 2), and 2D alignment and tracking (Table 5) benchmarks.

In-the-wild HMR. Table 1 compares Multi-HMR 2 with one-shot multi-person HMR methods on the 3DPW test set and EMDB1, without any finetuning on the 3DPW training set. Multi-HMR 2 significantly outperforms prior methods on both benchmarks, especially in terms of TA-PVE and Abs-PVE, which are the metrics that take into account the 3D localization in the camera space. AiOS particularly struggles to predict 3D locations because of its weak-perspective camera assumption. SAT-HMR also makes strong assumptions on the camera, but a fixed FOV of 60 degrees is much more realistic than a fixed focal length of 5000 mm. In particular, Multi-HMR 2 outperforms Multi-HMR in terms of 3D localization on 3DPW while using a predicted camera.

Interactions. We evaluate Multi-HMR 2 in the context of close interactions in Table 2. We achieve state-of-the-art results in pelvis-centered metrics while improving over the state of the art in interaction recovery and placement in the scene, even with a predicted camera. In the context of closely interacting humans, Multi-HMR 2 achieves a F1-score of 100% on Hi4D and detects 97% of annotated humans in Harmony4D, making it the strongest detector together with [53]. Table 3 shows that Multi-HMR 2 with a predicted camera significantly outperforms prior works for multi-person detection and translation estimation on the CMU-Panoptic [20] dataset.

Table 5: Reprojected keypoints (left) on MSCOCO [29] (‡ denotes methods using datasets with 2D annotations for training) and **tracking results** (right) on the PoseTrack21 [10] validation set.

Method	AP ↑	AP50 ↑	AP75 ↑	AR ↑	AR50 ↑	AR75 ↑
AiOS [54]	44.9	76.0	46.4	54.0	**83.3**	57.7
SAT-HMR‡ [53]	42.1	73.8	43.8	49.5	78.4	53.0
Multi-HMR [3]	31.2	60.0	29.1	39.0	66.0	39.8
Multi-HMR 2.a	34.0	61.2	34.4	39.3	63.4	42.1
Multi-HMR 2.b‡	**50.7**	**81.0**	**54.3**	**56.2**	83.2	**60.9**

Method	IDF1 ↑	MOTA ↑
Trained on HMR images		
Humans4D [15]	72.5	60.2
Multi-HMR 2.b	**75.2**	**65.5**
Trained on HMR videos		
CoMotion [39]	**79.5**	**70.1**

Camera estimation. Overall, Multi-HMR 2 with predicted intrinsics outperforms Multi-HMR that uses the ground-truth camera, even on metrics in the camera space. This result is impressive, since our location predictions are obtained as an inverse projection using our predicted 2D location, depth, and focal length. To better understand this result, we evaluate the vertical FOV predictions on 3DPW and EMDB and compare it with FOV estimation methods in Table 4. We achieve results comparable to specialized FOV estimation methods, which is why using a predicted camera has a reasonable impact on our performance. Importantly, the predicted intrinsics are also temporally stable within each video: the per-video standard deviation of the predicted FOV is only 2.5°on 3DPW and 2.6°on EMDB on average. This stability is desirable for video-based human reconstruction, where camera parameters should remain consistent over time. In particular, while most methods are either accurate on 3DPW or EMDB, Multi-HMR 2 is among the best approaches on both benchmarks. Aside from the 3D location, predicted human meshes are not conditioned on the camera. Pelvis-centered metrics (PVE, MPJPE, and PA-MPJPE) should then have the same value in both contexts, as centering on the pelvis eliminates location variations. While we observe this result on EMDB, we have slight difference in other datasets: this is due to the evaluation matching process relying on the reprojected joints.

Detection. Table 5 reports results on the MSCOCO val2017 set, using the MSCOCO API [29]. Note that when not given camera intrinsic parameters, Multi-HMR adopts a fixed FOV of 60 degrees. Multi-HMR 2 obtains the best re-projections, both in terms of precision and recall. Again, note that Multi-HMR 2 relies on a predicted camera while others use fixed or simplified camera models. Methods with learned perspective cameras face additional difficulty: the 3D pose must be consistent with varying cameras in order to reach high 2D precision. Despite this disadvantage, we demonstrate that jointly reasoning about detection, meshes, and camera leads to more reliable high-confidence predictions.

Tracking. We report the IDF1 and MOTA metrics on the PoseTrack21 validation set in Table 5. Multi-HMR 2 outperforms Humans4D, which also recovers meshes frame by frame and associate the detections over frames. In their case, the meshes are recovered with a two-stage framework (detection then single-person mesh estimation) and the association is done with an Hungarian matching on cost matrix combining the distances between predicted and estimated pelvis (as

Table 6: Ablation studies. (Top left) **Ablation on the training data.** S stands for synthetic data, R for real images with pseudo-ground truth meshes, and 2D for datasets with 2D annotations. (Bottom left) **Ablation study of the matching strategy.** (Top right) **Square padding at training and inference time.** (Bottom right) **Ablation on the predicted camera parameters.**

Data	3DPW		Hi4D		MSCOCO	
	PVE	PA-MPJPE	MPJPE	Pair-PA-MPJPE	AP	AR
S	93.6	48.2	64.4	**64.0**	11.9	24.7
R	**88.7**	45.7	67.9	75.7	21.3	26.6
S+2D	94.1	68.5	107.0	109.5	12.4	26.2
R+2D	94.5	50.1	105.9	113.9	20.6	**33.8**
S+R	90.5	**45.6**	**60.0**	77.8	**21.9**	29.4
S+R+2D	91.6	46.9	76.5	84.2	16.8	32.9

Matching	3DPW		Hi4D		MSCOCO	
	PVE	PA-MPJPE	MPJPE	Pair-PA-MPJPE	AP	AR
2D loc	**90.5**	**45.6**	**60.0**	77.8	**21.9**	29.4
3D loc	93.0	46.8	65.4	**76.4**	17.0	**30.4**
J3D	104.5	55.1	83.1	84.8	8.6	16.7

Train padding	Test padding	3DPW		Hi4D		MSCOCO	
		PVE	Abs-PVE	Abs-PVE	Pair-PA-MPJPE	AP	AR
✓	✓	90.5	45.6	60.0	77.8	21.9	29.4
✓	×	113.4	57.9	62.2	91.6	0.5	3.8
×	✓	89.7	46.3	61.7	77.8	19.8	32.2
×	×	85.8	43.8	61.5	74.7	17.7	30.9

Train camera	Test camera	3DPW		Hi4D		MSCOCO	
		PVE	Abs-PVE	Abs-PVE	Pair-PA-MPJPE	AP	AR
GT	GT	90.5	243.7	**148.0**	77.8	–	–
GT	pred	92.4	609.8	213.7	91.0	**21.9**	29.4
pred	pred	**86.2**	577.3	220.0	**72.7**	19.6	**31.0**
pred	GT	89.0	**243.4**	159.5	89.7	–	–

ours), the similarity between some handcrafted appearance features and finally the distances between some predicted poses and estimated pose. The latter pose prediction requires human motion sequences, in contrast to our approach that does not leverage any sequential information. We also report the performance of the recent CoMotion [39] approach that is trained to predict directly short HMR tracks on video clips, at the cost of a long stepwise training.

4.2 Ablation study

We ablate the main components of Multi-HMR 2. Ablation experiments use images of size 672, a ViT-L backbone initialized with DINOv2 [40], and the model is trained for 500k iterations with synthetic datasets and real images with pseudo-ground truth meshes. Additional ablations are in the supplementary material.

Training data. We first investigate the role of training data. We denote **R** the real images with pseudo-ground truth meshes from [41], **S** the synthetic datasets (AnnyOne and BEDLAM), and **2D** images from OpenImages [26] pseudo-labeled with ViTPose [64] estimations. Results in Table 6 show that on HMR benchmarks, **S** and **R** seem to benefit the most to mesh recovery overall, while using **2D** seems to degrade the 3D performance. This could be expected, as improvements in 2D and 3D performance often compete — improving one can lead to a decline in the other. When evaluating on MSCOCO, we realize that **S** is the most important to improve the precision, while **2D** improves significantly the recall by improving detections for individuals far away from the camera.

Matching strategy. We test different matching strategies (Table 6) with a fixed training budget, relying either on 2D localization in pixels, on 3D location,

or on 3D keypoints in the camera frame. While matching on 2D and 3D locations can be done without forwarding all queries to the parametric body model, 3D keypoints are more costly to obtain. Matching on 2D or 3D locations gives similar results, and matching on 3D joints gives much larger errors in all metrics. Multi-HMR 2 uses 2D matching as it allows us to use training data without 3D annotations.

Varying aspect ratios. We experiment using different aspect ratios at training, and evaluate the impact of square padding at inference, in Table 6. When training only with squared images, the performance decrease significantly when testing with diverse aspect ratios: this is because the model never saw this kind of images during training. While using squared images gives the best 2D performance, training and testing with diverse aspect ratios gives the best 3D metrics. We note that when training with diverse aspect ratios, testing with squared images has little impact on the performance.

Predicted camera. We experiment using ground-truth or predicted camera intrinsic parameters at train and test times (Table 6). While the ground-truth camera is consistently better to estimate the location in the camera coordinates system, other metrics – including the Pair-PA-MPJPE that measures interaction recoveries – benefit from using the same settings during training and inference.

5 Conclusion

We have introduced Multi-HMR 2, a robust DETR-based multi-person HMR method leveraging the Anny parametric body model. It outputs strong detections, accurate localization with a predicted camera, precise human mesh recovery in scene- and pelvis-centered coordinate systems, and features for tracking. Even if Multi-HMR 2 can be trained on a single GPU in about a week, future works may explore more advanced DETR strategies [8, 17, 32] for faster convergence. Other lines of improvement include modeling camera distortions [58,59] to be more robust to images captured with extreme camera angles such as selfies.

References

1. Almazan, J., Gajic, B., Murray, N., Larlus, D.: Re-id done right: towards good practices for person re-identification. arXiv preprint arXiv:1801.05339 (2018)
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. Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., Lucas, T.: Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In: ECCV (2024)
4. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR (2023)
5. Brégier, R., Fiche, G., Bravo-Sánchez, L., Lucas, T., Armando, M., Weinzaepfel, P., Rogez, G., Baradel, F.: Human mesh modeling for Anny body. In: ECCV (2026)

6. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: SMPLer-X: Scaling up expressive human pose and shape estimation. In: NeurIPS (2023)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
8. Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G., Wang, J.: Group DETR: Fast detr training with group-wise one-to-many assignment. In: ICCV (2023)
9. 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)
10. Doering, A., Chen, D., Zhang, S., Schiele, B., Gall, J.: PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In: CVPR (2022)
11. 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)
12. Facil, J.M., Ummenhofer, B., Zhou, H., Montesano, L., Brox, T., Civera, J.: CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In: CVPR (2019)
13. Fulgeri, F., Fabbri, M., Alletto, S., Calderara, S., Cucchiara, R.: Can adversarial networks hallucinate occluded people with a plausible aspect? CVIU (2019)
14. Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., Tran, D.: Detect-and-track: Efficient pose estimation in videos. In: CVPR (2018)
15. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: ICCV (2023)
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
17. Huang, S., Lu, Z., Cun, X., Yu, Y., Zhou, X., Shen, X.: DEIM: Detr with improved matching for fast convergence. In: CVPR (2025)
18. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: CVPR (2020)
19. Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Blackburn-Matzen, K., Sticha, M., Fouhey, D.F.: Perspective fields for single image camera calibration. In: CVPR (2023)
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. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
22. 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)
23. Khirodkar, R., Song, J.T., Cao, J., Luo, Z., Kitani, K.: Harmony4d: A video dataset for in-the-wild close human interactions. In: NeurIPS (2024)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
25. 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)
26. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020)

27. Lee, J., Go, H., Lee, H., Cho, S., Sung, M., Kim, J.: Ctrl-C: Camera calibration transformer with line-classification. In: ICCV (2021)
28. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
29. 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)
30. Liu, H., Chen, Q., Tan, Z., Liu, J.J., Wang, J., Su, X., Li, X., Yao, K., Han, J., Ding, E., et al.: Group pose: A simple baseline for end-to-end multi-person pose estimation. In: ICCV (2023)
31. Liu, K., Fu, Y., Yuan, W., Lin, J., Li, P., Gu, X., Qiu, L., Wang, H., Dong, Z., Han, X.: Motions as queries: One-stage multi-person holistic human motion capture. In: CVPR (2025)
32. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: ICLR (2022)
33. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: ECCV (2016)
34. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (2015)
35. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE trans. Multimedia (2019)
36. Luvizon, D.C., Habermann, M., Golyanik, V., Kortylewski, A., Theobalt, C.: Scene-aware 3d multi-human motion capture from a single camera. In: Computer Graphics Forum (2023)
37. 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)
38. Mertan, A., Duff, D.J., Unal, G.: Single image depth estimation: An overview. Digital Signal Processing (2022)
39. Newell, A., Hu, P., Lipson, L., Richter, S.R., Koltun, V.: CoMotion: Concurrent multi-person 3d motion. In: ICLR (2025)
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., Black, M.J.: CameraHMR: Aligning people with perspective. In: 3DV (2025)
42. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.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: ACM MM (2022)
45. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people with 3d representations. In: NeurIPS (2021)
46. Rajasegaran, J., Pavlakos, G., Kanazawa, A., Malik, J.: Tracking people by predicting 3D appearance, location and pose. In: CVPR (2022)

47. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE trans. PAMI (2020)
48. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
49. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, real-time object detection. In: CVPR (2016)
50. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al.: Hiera: A hierarchical vision transformer without the bells-and-whistles. In: ICML (2023)
51. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: CVPR (2022)
52. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3. TMLR (2026)
53. Su, C., Ma, X., Su, J., Wang, Y.: SAT-HMR: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In: CVPR (2025)
54. Sun, Q., Wang, Y., Zeng, A., Yin, W., Wei, C., Wang, W., Mei, H., Leung, C.S., Liu, Z., Yang, L., et al.: AiOS: All-in-one-stage expressive human pose and shape estimation. In: CVPR (2024)
55. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV (2021)
56. 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)
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
58. Wang, S., Li, J., Li, T., Yuan, Y., Fuchs, H., Nagano, K., De Mello, S., Stengel, M.: BLADE: Single-view body mesh estimation through accurate depth estimation. In: CVPR (2025)
59. Wang, W., Ge, Y., Mei, H., Cai, Z., Sun, Q., Wang, Y., Shen, C., Yang, L., Komura, T.: Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In: ICCV (2023)
60. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: PromptHMR: Promptable human mesh recovery. In: CVPR (2025)
61. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019), Accessed: 2026-06-25
62. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018)
63. 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)
64. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: NeurIPS (2022)
65. Yang, J., Zeng, A., Liu, S., Li, F., Zhang, R., Zhang, L.: Explicit box detection unifies end-to-end multi-person pose estimation. In: ICLR (2023)
66. Yin, Y., Guo, C., Kaufmann, M., Zarate, J., Song, J., Hilliges, O.: Hi4D: 4d instance segmentation of close human interaction. In: CVPR (2023)
67. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR (2018)

68. 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)
69. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
70. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR (2018)
71. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: arXiv preprint arXiv:1904.07850 (2019)
72. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR (2019)
73. Zhu, S., Kumar, A., Hu, M., Liu, X.: Tame a wild camera: In-the-wild monocular camera calibration. In: NeurIPS (2023)
74. Zhu, Y., Li, A., Tang, Y., Zhao, W., Zhou, J., Lu, J.: Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In: CVPR (2024)