MASK-HMR: REAL-TIME MULTI-PERSON HUMAN MESH RECOVERY WITH IMPLICIT MASK PROMPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masks provide informative visual prompts for human mesh recovery (HMR). However, existing methods using mask prompts for HMR typically rely on an upstream segmentation model, increasing computational cost and inference latency. We propose Mask-HMR, a one-stage multi-person human mesh recovery framework with implicit mask prompts. Mask-HMR follows a DEtection TRansformer (DETR) pipeline, in which a fixed set of learnable human queries is initialized and decoded into mesh parameters. In Mask-HMR, each human query learns to encode implicit instance and body-part masks of the associated human instance. A learned generator maps the query to a pointwise mask prediction head, which predicts instance and body-part masks of the associated human at a given image point. By encoding mask information directly in the human queries, Mask-HMR benefits from mask prompts without dense mask prediction. In addition, instead of full mask supervision, the implicit masks and the generator can be trained with a small set of sampled points. Mask-HMR thus gains the benefit of mask prompting with limited inference and training costs. We also design a per-joint pose refinement module at the final layer, where the pointwise mask prediction head of each query is evaluated in the region of interest to extract joint-relevant features for joint pose refinement. Experiments show that Mask-HMR achieves state-of-the-art performance among one-stage multi-person HMR methods on both standard and close-interaction benchmarks while retaining real-time inference.

1 INTRODUCTION

Multi-person human mesh recovery (HMR) from a single RGB image is a fundamental task with broad applications, including healthcare, human-robot interaction, and animation (Stenum et al., 2024; Svenstrup et al., 2009; Xia et al., 2017). Earlier methods follow a two-stage pipeline, an upstream model first detects and crops each human instance (Ravi et al., 2025; Yang et al., 2026) and a separate model then recovers a human mesh from each crop (Yang et al., 2026). Although these approaches have shown promising results, they depend on upstream detection models and introduce additional inference latency. Moreover, cropping each person independently removes global image context, which can be important for resolving ambiguous 3D pose configurations.

Single-stage methods recover all human meshes from the image in one forward pass. Multi-HMR represents each person using the image token at a primary keypoint, such as the head (Baradel et al., 2025). Each query then recovers the mesh of the associated human instance. However, the model fails when the primary keypoint is occluded. Another line of work adopts the Detection Transformer (DETR) framework, in which a fixed set of learnable queries attend to the global image features through a transformer decoder. After the decoder, each query produces a candidate human mesh with an associated confidence score indicating whether the query corresponds to a person. In previous works, AiOS and SAT-HMR adopt the DETR structure and associate human queries with anchor boxes (Sun et al., 2024; Su et al., 2025). Although spatial cues such as anchor boxes localize people efficiently and accelerate training, they can become ambiguous when people interact closely.

Human mesh recovery and segmentation are closely related. An instance mask provides a 2D silhouette constraint for body mesh recovery and can serve as a prompt for distinguishing human instances, even under occlusion or overlap (Wang et al., 2025; Yang et al., 2026). Body-part masks can also guide mesh recovery. For example, PARE (Kocabas et al., 2021) extracts the feature associated with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

each joint and uses the extracted features to predict the parameters of that joint. However, these methods typically require an additional segmentation model or structure, which increases computational cost and inference latency (Wang et al., 2025; Yang et al., 2026).

In this paper, we introduce Mask-HMR, a one-stage DETR framework that recovers multiple 3D human meshes with implicit mask prompts. In Mask-HMR, each query learns to encode implicit instance and body-part masks of the associated person. Through a learned generator, the query can be mapped to a pointwise mask prediction head, which predicts instance and body-part masks of the associated human instance at a given point in the image. The query thus encodes an implicit mask prompt for mesh recovery without requiring additional segmentation models. In addition, instead of full mask supervision, the implicit masks and the generator can be trained with a small set of sampled points. Mask-HMR thus gains the benefit of mask prompting with limited inference and training costs. We also design a per-joint pose refinement module in which the pointwise mask prediction head of each query is evaluated at the final layer to extract joint-relevant features for the refinement of each joint pose. We evaluate Mask-HMR on the standard 3DPW (von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023) benchmarks and the human interaction benchmarks Hi4D (Yin et al., 2023) and Harmony4D (Khirodkar et al., 2024). Mask-HMR outperforms prior single-stage multi-person HMR methods while retaining real-time inference, and qualitative results on in-the-wild images demonstrate its generalization.

Our main contributions are as follows:

1. We propose Mask-HMR, a one-stage framework that recovers all human meshes in an RGB image. The model is also capable of producing instance and body-part segmentations.
2. We introduce implicit mask prompts. In Mask-HMR, each query learns to encode the instance and body-part masks of the associated human instance implicitly. With the implicit mask design, Mask-HMR gains the benefit of mask prompting with limited training and inference costs. We also design a per-joint pose refinement module, where the masks are evaluated explicitly to guide the joint-relevant feature extraction and refine the final joint estimation.
3. We show that Mask-HMR achieves state-of-the-art performance among one-stage multi-person HMR methods on both standard and human interaction benchmarks while retaining real-time inference.

2 Related Work

2.1 Multi-person human mesh recovery

Multi-person human mesh recovery aims to estimate the parameters of a parametric body model (Loper et al., 2015; Pavlakos et al., 2019) for every person in an image. Early approaches adopt a two-stage pipeline: a detector first crops each person, after which a mesh recovery model processes every crop independently (Yang et al., 2026; Shen et al., 2024). However, the two-stage pipeline leads to higher inference latency. In addition, cropped images lack global image context, making it challenging to handle occlusion (Sun et al., 2021).

Single-stage methods instead recover all people in a single forward pass. Multi-HMR (Baradel et al., 2025) first detects a primary keypoint for each human instance, such as the head, and uses the corresponding image token as a query for whole-body mesh recovery. Although efficient, this approach may miss occluded people whose primary keypoints are not visible. Another line of work builds on the DEtection TRansformer (DETR) (Carion et al., 2020), representing each person with a query that is decoded into parametric body model parameters. AiOS and SAT-HMR associate each human query with an anchor box that guides attention toward the region of the corresponding person (Liu et al., 2022; Sun et al., 2024; Su et al., 2025). However, a bounding box can be ambiguous when the human instances closely interact, and different queries may attend to similar features.

Mask-HMR adopts a DETR-style architecture with implicit masks embedded in each human query. The masks provide fine-grained spatial guidance for attention and separate people even when they closely interact. Moreover, a human mask closely approximates the 2D projection of the corresponding mesh. The mask supervision thus constrains mesh recovery, accelerating convergence and improving accuracy.

2.2 MASKS FOR HUMAN MESH RECOVERY

Previous work has explored instance and body-part masks as prompts for human mesh recovery. PromptHMR and SAM 3D Body use instance masks as prompts for mesh recovery (Wang et al., 2025; Yang et al., 2026), improving estimation accuracy and robustness when people overlap. PARE (Kocabas et al., 2021) predicts body-part masks and uses the relevant regions to guide joint-specific mesh estimation, improving accuracy and robustness under occlusion.

Although these approaches demonstrate the value of masks for mesh recovery, they rely on an upstream segmentation model, such as Segment Anything Model 2 (SAM 2) (Ravi et al., 2025), which introduces additional inference cost. To address this limitation, we view human mesh recovery and segmentation as an integrated task and train the model to directly embed the masks into the queries. Queries carrying the mask information thus provide the mask prompt without an additional segmentation model or structure. Inspired by PARE, we also design a per-joint pose refinement module. The masks are evaluated in the person-centric region and guide the extraction of the joint-relevant features. The features are then used to refine the pose of that specific joint. For example, the feature patches of the upper arm and the forearm can be extracted and used to refine the elbow pose. With these designs, Mask-HMR is able to achieve state-of-the-art performance among one-stage multi-person human mesh recovery methods while retaining real-time inference.

2.3 IMPLICIT MASKS WITH POINTWISE SUPERVISION

Segmentation models are conventionally trained with full supervision against every pixel in the ground-truth masks, which is computationally demanding. PointRend (Kirillov et al., 2020) reduces this cost by computing the mask loss only at a small set of points sampled from uncertain regions. Implicit PointRend (Cheng et al., 2022) extends this idea by representing each instance mask as a pointwise mask prediction function that maps image coordinates to foreground mask probabilities. Using only a small set of labeled points per object, Implicit PointRend matches the segmentation performance of fully supervised methods.

Directly embedding the mask into the query would require each query to learn to generate a full mask at the image resolution, which could be expensive for training. Instead, in Mask-HMR, each query learns to embed an implicit mask, which a learned generator maps to a pointwise mask prediction head. With this design, each query carries the complete mask information of the associated person, while the training can be done with a small set of sampled points in the image. As a result, in Mask-HMR, queries learn to embed the masks with limited training cost.

3 METHOD

Our model recovers all human meshes in the RGB image in one stage. The model follows a DEtection TRansformer (DETR) (Carion et al., 2020) pipeline. Given an input RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with resolution $H \times W$, the model initializes a fixed set of learnable human queries $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_q}$, where N_q is a hyperparameter that represents the maximum number of human instances that can be estimated. For each of the N_m human instances in the image, our model, denoted by $\mathcal{H}$, predicts a human mesh $\mathbf{M}_i \in \mathbb{R}^{V \times 3}$ composed of V vertices and a root translation $\mathbf{t}_i \in \mathbb{R}^3$ in the camera coordinate system. In addition, each query also carries the instance and body-part masks. Optionally, a query can be mapped to a pointwise mask prediction function to generate the instance and body-part masks $\mathbf{S}_i$ of the associated human instance (Section 3.2.3):

$$\{(\mathbf{M}_i, \mathbf{t}_i, [\mathbf{S}_i])\}_{i=1}^{N_m} = \mathcal{H}(\mathbf{I}, \mathbf{Q}). \tag{1}$$

Section 3.1 introduces the 3D parametric human mesh model, Section 3.2 presents the model architecture, and Section 3.3 describes the training losses.

3.1 HUMAN MESH MODEL REPRESENTATION

We use the Skinned Multi-Person Linear Model (SMPL) (Loper et al., 2015) as our human mesh model. Given joint parameters $\boldsymbol{\theta} \in \mathbb{R}^{24 \times 3}$ in axis-angle representation and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ that describe the body shape, SMPL outputs a 3D human mesh $\mathbf{M} = \text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{V \times 3}$

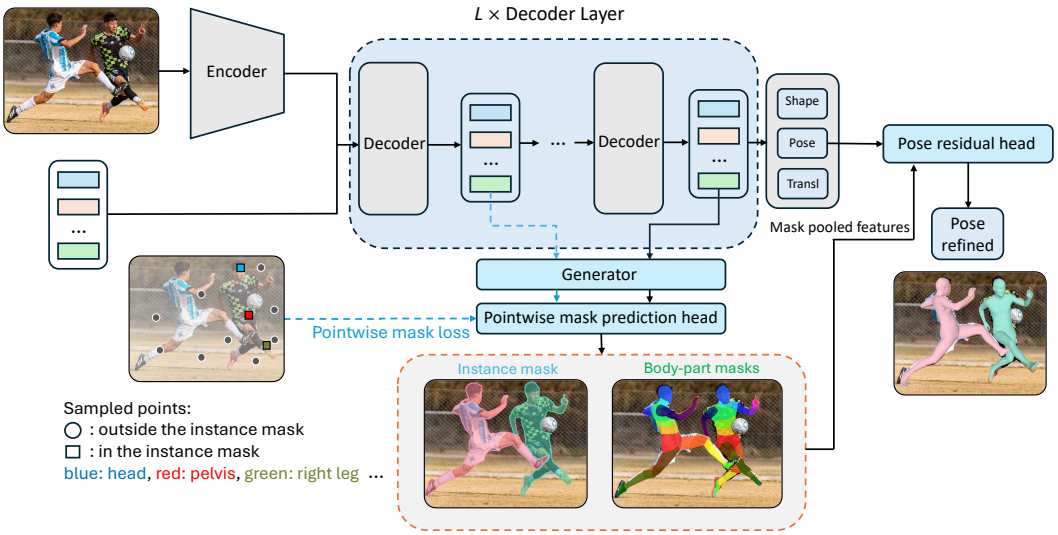

Figure 1: Architecture overview of Mask-HMR. Mask-HMR follows a DETR pipeline. Image tokens are extracted using a pretrained ViT encoder. A fixed set of queries is initialized and decoded into mesh parameters. Each human query additionally carries implicit masks, which can be mapped into a pointwise mask prediction head through a generator. The mask loss is computed at a small set of sampled points to supervise the generator and the implicit mask embeddings. In the final layer, the masks are evaluated at the region of interest and guide the per-joint pose refinement.

with $V = 6890$ vertices. The 3D joint locations are computed as $\mathcal{J}_{3D} = \mathbf{WM} \in \mathbb{R}^{J\times 3}$, where $J = 24$ is the number of whole-body joints and $\mathbf{W}$ is a pretrained linear regressor. Our model regresses the joint rotations in the 6D representation (Zhou et al., 2019), denoted $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{24\times 6}$. These 6D rotations are converted to the axis-angle form before being passed to SMPL.

3.2 Architecture

The overall framework of Mask-HMR is depicted in Figure 1. Given an input image, a ViT encoder (Section 3.2.1) extracts the image tokens. A fixed set of learnable human queries is initialized and decoded into human meshes through a DETR-style decoder (Section 3.2.2). Each decoded query represents one candidate human instance and predicts the mesh parameters, camera translation, and detection confidence. The query also carries implicit masks that a learned generator converts into a pointwise mask prediction head (Section 3.2.3).

Implicit masks. Each decoded human query carries implicit instance and body-part masks of the associated human instance. During training, after each decoder layer, a learned generator maps every matched query to the parameters of a pointwise mask prediction head. We supervise the predictions of each generated head at sampled image points, thereby training the generator and encouraging the query to embed the mask information. The embedded implicit masks serve as the mask prompt for the mesh recovery without dense mask prediction. During inference, we generate and evaluate the pointwise mask prediction head from the final-layer queries, and the resulting masks are used by the per-joint pose refinement module to refine each joint pose estimation (Section 3.2.4).

3.2.1 Encoder

The encoder uses a Vision Transformer (ViT) backbone pretrained with DINOv2 (Oquab et al., 2024). Given the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the encoder divides it into patches of size $P \times P$, each embedded into tokens with a linear transformation and positional embeddings. The set of tokens is then processed with self-attention blocks into image tokens $\mathbf{E} \in \mathbb{R}^{H/P\times W/P\times D}$, where D is the feature dimension. The image tokens are then used as the cross-attention memory for the decoder.

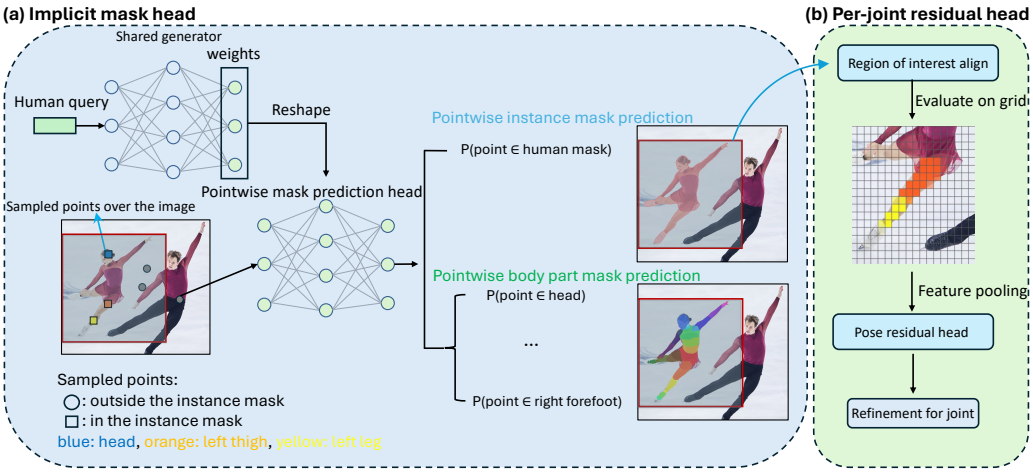

Figure 2: Implicit mask head (left) and per-joint residual head (right). Left: a learned generator maps each human query to the parameters of a pointwise mask prediction head. At every sampled point, this head combines interpolated image features with a positional encoding and predicts the instance-mask probability and body-part probabilities. Right: the pointwise mask prediction head is evaluated on a region-of-interest (RoI)-aligned grid within the person-centric region. The predicted instance and body-part probabilities weight the pooling of grid features into per-joint features, from which the pose residual head predicts a refinement for each joint.

3.2.2 Decoder

The decoder comprises six transformer layers that decode a fixed set of learnable human queries $\mathbf{Q} = \{\mathbf{q}_i\}_{i=1}^{N_q}$. Specifically, each decoder layer applies cross-attention between the human queries and the image tokens $\mathbf{E}$, self-attention over the human queries, and a feed-forward block. Between layers, a learned generator maps the updated queries to the parameters of a pointwise mask prediction head and is supervised against the sampled points in the image (Section 3.2.3). After each layer, MLP heads read each updated query and predict (i) the SMPL pose in the 6D rotation representation and the shape parameters, (ii) the camera parameters (Appendix B), and (iii) a confidence score indicating whether the query represents a person.

3.2.3 Implicit mask head

To inject segmentation information into each human query, we train the query to carry implicit masks. A learned generator maps the query to a pointwise mask prediction head. Given an image point, the pointwise mask prediction head jointly predicts (i) the instance-mask logit and (ii) the body-part mask logits. The full pipeline is shown in Figure 2.

Specifically, given the human query $\mathbf{q}_i$, the generator, a multi-layer perceptron $\mathrm{MLP}_{\mathrm{gen}}$, predicts the parameters of the pointwise mask prediction head for the associated person: $\mathbf{\Theta}_i = \mathrm{MLP}_{\mathrm{gen}}(\mathbf{q}_i)$. These parameters are then reshaped to form the pointwise mask prediction head $\mathrm{MLP}_{\mathbf{\Theta}_i}$. Given a sampled image point $\mathbf{p}_k$ and its feature $\mathbf{x}_k$, this pointwise mask prediction head predicts an instance-mask logit and body-part logits:

$$\left[\hat{m}_{i,k},\, \hat{\mathbf{s}}_{i,k}\right] = \mathrm{MLP}_{\mathbf{\Theta}_i}\big([\mathbf{x}_k,\, \mathrm{PE}(\mathbf{p}_k)]\big), \tag{2}$$

where $\hat{m}_{i,k} \in \mathbb{R}$ is the instance mask logit at $\mathbf{p}_k$, and $\hat{\mathbf{s}}_{i,k} \in \mathbb{R}^{N_P+1}$ contains body-part logits for the $N_P = 24$ SMPL body parts and a null class for foreground pixels that do not correspond to an SMPL body part, such as pixels depicting loose clothing or hair. $\mathrm{PE}(\mathbf{p}_k)$ denotes the positional encoding of the sampled point. The ViT tokens are computed at patch resolution, which is too coarse to capture high-frequency details such as mask boundaries. We therefore pair the ViT with a lightweight convolutional neural network (CNN) branch that operates at a higher spatial resolution, i.e., $\mathbf{x}_k = \left[\mathbf{x}^{\mathrm{vit}}(\mathbf{p}_k),\, \mathbf{x}^{\mathrm{cnn}}(\mathbf{p}_k)\right]$ concatenates features bilinearly sampled at image location $\mathbf{p}_k$ from the two branches.

Pointwise supervision. For each image, we randomly sample a set of points. We then apply the mask losses to the predictions of every pointwise mask prediction head at these points. To provide denser supervision for body-part mask prediction, we sample points at five times the background density around each person, defined by the bounding box of the ground-truth 2D joints.

3.2.4 PER-JOINT POSE REFINEMENT

After the final decoder layer, each query $\mathbf{q}_i$ produces an initial pose estimation $\hat{\boldsymbol{\theta}}_i^{(0)}$, camera parameters $\mathbf{c}_i$, and the parameters $\boldsymbol{\Theta}_i$ of a pointwise mask prediction head $\mathrm{MLP}_{\boldsymbol{\Theta}_i}$ of the associated human instance. We refine the initial pose using features pooled according to the predicted instance and body-part masks. The refinement head predicts a rotation residual $\Delta\hat{\boldsymbol{\theta}}_{i,j}$ for each joint j.

Body-part aware feature pooling. Around each person, we define a square region $\mathcal{R}_i$ whose side length is determined by the predicted camera scale s_i. Intuitively, $\mathcal{R}_i$ corresponds to the image region of a 2 m square placed at the depth of the instance (Appendix B). Region of Interest (RoI) Align then maps the ViT features within $\mathcal{R}_i$ to a $G \times G$ grid of $N_r = G^2$ cells. We evaluate the pointwise mask prediction head at the center of each grid cell to obtain the corresponding instance logit $\hat{m}_{i,k}$ and body-part logits $\hat{\mathbf{s}}_{i,k}$.

Let $\hat{\pi}_{i,p,k} = \mathrm{softmax}(\hat{\mathbf{s}}_{i,k})_p$ denote the predicted probability that cell k belongs to body part p. Combining it with the instance mask probability, we define the part visibility score $w_{i,p,k} = \hat{\pi}_{i,p,k}\,\sigma(\hat{m}_{i,k})$, where $\sigma(\cdot)$ is the sigmoid function that maps the instance mask logit to a probability. Intuitively, $w_{i,p,k}$ is the probability that body part p of human instance i is visible at cell k. We normalize the visibility scores across the N_r cells together with a learned abstention term $\exp(a_p)$ to obtain the pooling weights. The abstention term allows the model to extract no features when that part is completely occluded.

$$\alpha_{i,p,k} = \frac{w_{i,p,k}}{\exp(a_p) + \sum_{k'=1}^{N_r} w_{i,p,k'}}, \qquad k \in \{1, \ldots, N_r\}. \tag{3}$$

We then compute each body-part feature as a weighted sum of the ViT features:

$$\mathbf{f}_{i,p} = \sum_{k=1}^{N_r} \alpha_{i,p,k}\, \mathbf{x}_{i,k}^{\mathrm{vit}}. \tag{4}$$

Joint feature construction. Estimating the articulation of a joint requires information from the body parts adjacent to that joint. Let $p_1(j)$ and $p_2(j)$ denote the two body parts associated with joint j, ordered from proximal to distal along the kinematic chain. We define the pooled feature for a joint as the concatenation of the features of the associated body parts, $\mathbf{f}_{i,j} = \left[\mathbf{f}_{i,p_1(j)}, \mathbf{f}_{i,p_2(j)}\right]$. For example, the feature for the elbow joint combines the upper-arm and forearm features. For the pelvis, which lies within a single body part, we duplicate its pooled feature in both entries. This construction yields $J = 24$ features $\{\mathbf{f}_{i,j}\}_{j=1}^{J}$ for each joint.

Residual prediction. A shared MLP predicts the refinement of each joint from the pooled features. The input to the MLP is $\mathbf{z}_{i,j} = \left[\mathbf{f}_{i,j},\, \mathbf{q}_i,\, \hat{\boldsymbol{\theta}}_i^{(0)},\, \mathbf{e}_j\right]$, which combines the features related to the joint $\mathbf{f}_{i,j}$, the decoder query $\mathbf{q}_i$, the initial pose estimation $\hat{\boldsymbol{\theta}}_i^{(0)}$, and a learned joint-identity embedding $\mathbf{e}_j$. The joint-identity embedding allows the shared head to specialize its prediction for each joint. The residual head predicts $\Delta\hat{\boldsymbol{\theta}}_{i,j} = \mathrm{MLP}_{\mathrm{res}}(\mathbf{z}_{i,j})$, and the final rotation of joint j is $\hat{\boldsymbol{\theta}}_{i,j} = \hat{\boldsymbol{\theta}}_{i,j}^{(0)} + \Delta\hat{\boldsymbol{\theta}}_{i,j}$.

3.3 TRAINING LOSSES

Following DETR (Carion et al., 2020), we assign predicted queries to ground-truth instances by bipartite matching with the Hungarian algorithm. The matching cost combines the query confidence, instance-mask, and projected 2D joint costs. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{conf}} + \mathcal{L}_{\mathrm{mesh}} + \mathcal{L}_{\mathrm{mask}}. \tag{5}$$

The confidence loss $\mathcal{L}_{\mathrm{conf}}$ supervises all queries, while the mesh and mask losses are evaluated only over matched pairs. The mesh-recovery loss $\mathcal{L}_{\mathrm{mesh}}$ is a weighted combination of losses on 3D joints,

projected 2D joints, SMPL pose and shape parameters, SMPL vertices, and root depth. The mask loss $\mathcal{L}_{\text{mask}}$ is evaluated at $K = 300$ sampled image points shared by all queries. The mask loss comprises cross-entropy and Dice terms for the instance and body-part masks. Additional details are provided in Appendix C.1.

4 EXPERIMENTS

4.1 DATASETS

4.1.1 TRAINING

We train Mask-HMR on a mix of standard human mesh recovery datasets: BEDLAM (Black et al., 2023), AGORA (Patel et al., 2021), COCO (Lin et al., 2014), MPII (Andriluka et al., 2014), AIC (Wu et al., 2019), SA-1B (Kirillov et al., 2023), and MOYO (Tripathi et al., 2023). BEDLAM and AGORA are synthetic datasets that natively provide 3D mesh and instance-mask annotations. For COCO, MPII, AIC, and SA-1B, we use meshes and masks from SAM 3D Body (Yang et al., 2026). In the SAM 3D Body annotations, some human instances in a subset of images are left unannotated, so we filter out those images before training. The dataset mix results in approximately 400k images. For MOYO, we generate pseudo instance masks with SAM 3 (Carion et al., 2026). Ground-truth body-part masks are obtained by projecting the SMPL mesh onto the image plane and transferring its per-face part labels (Loper et al., 2015) to the pixels within the corresponding human-instance mask. Pixels within the instance mask but do not receive a projected body-part label are assigned to the null class. Further details about training-data processing are provided in Appendix A.

4.1.2 EVALUATION

For comparison with state-of-the-art methods, we evaluate on 3DPW (von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023). To assess performance during close interactions between two people, we also evaluate our method on the Hi4D and Harmony4D datasets. Following previous work (Baradel et al., 2025; Sun et al., 2021), we report mean per-joint position error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), mean vertex error (MVE), and normalized mean vertex error (NMVE). For the interaction datasets Hi4D and Harmony4D, we also report Pair-PA-MPJPE (Pair-PA). Compared with PA-MPJPE, Pair-PA additionally measures the errors in the relative position and orientation between the two people (Müller et al., 2024). All metrics above are reported in millimeters (mm). To evaluate computational cost, we report the average inference time (ms) on a single RTX 4090 GPU. A more detailed description of the metrics is provided in Appendix G.

4.2 IMPLEMENTATION DETAILS

The input image is resized to maintain its original aspect ratio, with the longer side of the image set to 896 pixels. We use a DINOv2-pretrained ViT-B (Oquab et al., 2024) as the encoder and a six-layer transformer decoder with 30 human queries. To balance accuracy and computational cost, we sample 300 points in each image to supervise the mask predictions (Sec. D.1). We train the model for 60 epochs with a batch size of 48. Training takes approximately 10 days on four RTX 4090 GPUs. The initial learning rates are 2×10^{-5} for the encoder and 4×10^{-5} for the remaining parameters. Further optimization and loss details are provided in Appendix C.1.

4.3 COMPARISON WITH THE STATE OF THE ART

4.3.1 PERFORMANCE AND EFFICIENCY

We evaluate Mask-HMR on the 3DPW (von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023) datasets, as shown in Table 1. Mask-HMR achieves state-of-the-art performance, with markedly higher estimation accuracy than previous single-stage multi-person mesh recovery methods. Mask-HMR runs at 18.3 ms per image, a runtime comparable to that of previous real-time methods such as SAT-HMR (16.2 ms), demonstrating its potential for real-time applications.

For close-interaction reconstruction, Mask-HMR also achieves state-of-the-art accuracy, as shown in Table 2. Our method achieves the best overall accuracy on both per-person and inter-person metrics.

Table 1: Comparison of model performance on 3DPW and EMDB. Number of joints evaluated are in parentheses. "Res." denotes the input resolution. "644*" indicates that SAT-HMR uses a base resolution of 644 with adaptive scaling up to 1288. Inference time is measured on BEDLAM validation set using an RTX 4090 GPU. ViT backbones are pretrained with DINOv2. Reconstruction metrics are in mm. **Best** and second-best results in each column are **bolded** and underlined, respectively.

Model	Backbone	Res.	Inference times (ms) ↓	3DPW (14)			EMDB (24)		
				PA-MPJPE ↓	MPJPE ↓	MVE ↓	PA-MPJPE ↓	MPJPE ↓	MVE ↓
AiOS	ResNet-50	1333	169.3	49.0	75.2	89.5	62.6	96.0	114.0
BEV	HRNet-32	512	18.0	46.9	78.5	92.3	70.6	109.2	129.1
SAT-HMR	ViT-B	644*	**16.2**	52.7	81.0	94.6	70.8	117.9	131.9
Multi-HMR	ViT-L	896	48.6	46.9	69.5	88.8	50.3	82.1	95.9
Mask-HMR	ViT-B	896	18.3	**38.3**	**60.5**	**71.8**	**43.0**	**70.0**	**83.5**

Table 2: Comparison on the close-interaction datasets Hi4D and Harmony4D. Reconstruction metrics are in mm. **Best** and second-best results in each column are **bolded** and underlined, respectively.

Model	Hi4D (14)				Harmony4D (14)			
	PA-MPJPE ↓	MPJPE ↓	Pair-PA ↓	Recall ↑	PA-MPJPE ↓	MPJPE ↓	Pair-PA ↓	Recall ↑
AiOS	57.6	83.5	302.5	98.9	**54.5**	82.2	339.8	**97.5**
BEV	65.2	101.1	114.4	97.0	81.3	132.3	208.9	93.9
SAT-HMR	62.0	95.5	97.3	**99.4**	59.3	90.8	176.8	97.3
Multi-HMR	49.6	67.8	83.0	96.5	67.0	115.6	288.5	93.7
Mask-HMR	**38.9**	**56.0**	**80.8**	98.4	**54.5**	**78.6**	**173.4**	96.3

Additional evaluation on the segmentation performance of the model is presented in Appendix F.

4.3.2 Qualitative results

Figure 3 presents qualitative results of our approach. We visualize both the mask and mesh predictions. Mask-HMR performs well across a variety of poses and in crowded scenes.

Figure 5 provides additional qualitative comparisons of Mask-HMR with other multi-person mesh recovery methods. Our method delivers more accurate estimates than previous methods, especially under occlusion. We also provide failure cases of our method in Figure 6. Mask-HMR can fail in cases involving extreme poses and heavy occlusion. Mask-HMR also cannot prevent interpenetration between closely interacting people due to the inherent depth ambiguity in monocular images.

4.4 Ablations

For the ablations, we train our model on the downsampled BEDLAM (Black et al., 2023) and AGORA (Patel et al., 2021) datasets for 40 epochs and evaluate the model on their validation sets to verify the effectiveness of the proposed components and hyperparameters. We ablate (i) the effects of the implicit masks and per-joint pose residual head (Sec. 4.4.1) and (ii) the effects of the number of points sampled during training (Appendix D.1).

4.4.1 Effects of the implicit masks and per-joint pose residual head

As shown in Figure 4, we compare four variants of the model. At epoch 40, adding the implicit instance masks reduces NMVE by $3.6/3.5$ mm on the BEDLAM/AGORA validation sets, while adding the body-part masks yields a further $2.2/1.0$ mm reduction. These results verify that the implicit mask provides spatial and 2D silhouette priors that facilitate learning, while the body-part embeddings add strong anatomical priors and further improve performance. Compared with NMVE, PA-MPJPE focuses on errors in local joint articulation. Adding the per-joint pose residual module reduces PA-MPJPE by $1.6/1.4$ mm, confirming that the refinement module produces more accurate joint estimates. Overall, compared with the DETR baseline, Mask-HMR reduces PA-MPJPE by 6.6



Figure 3: Qualitative examples of Mask-HMR. For each example, we show (i) the input image, (ii) the predicted instance mask, (iii) the predicted body-part masks, (iv) the mesh prediction overlaid on the image, and (v) a 3D view of the mesh prediction.

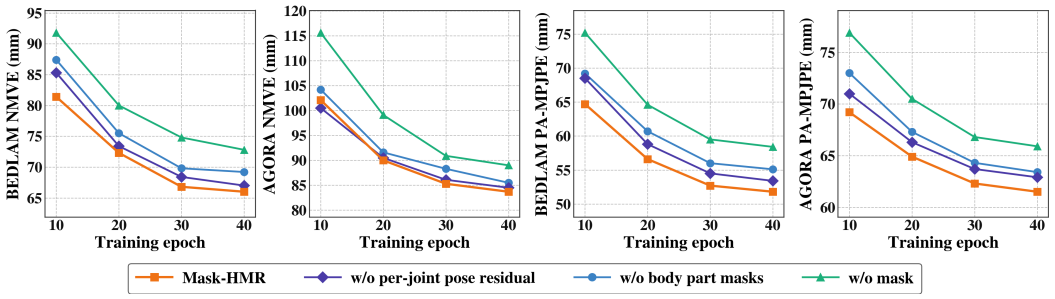

Figure 4: Ablation of the model components. We evaluate four model variants on the BEDLAM and AGORA validation sets: (i) Mask-HMR, (ii) the model without the per-joint pose residual module (w/o per-joint residual), (iii) the model with instance-level implicit masks but without body-part masks (w/o body-part masks), and (iv) a plain DETR baseline (w/o mask). Measured on a single RTX 4090, the four variants run at 18.3, 16.5, 16.4, and 14.7 ms per image, respectively.

and 4.4 mm and NMVE by 6.8 and 5.3 mm on BEDLAM and AGORA, validating the effectiveness of the proposed designs.

5 Conclusion

In this paper, we present a novel one-stage framework that recovers all human meshes from an RGB image. Instead of relying on an upstream segmentation model, each query embeds the instance and body-part masks implicitly. The design naturally integrates the human mesh recovery and segmentation tasks, enabling our method to improve mesh recovery accuracy using mask prompts while maintaining real-time inference. Experiments show that our model establishes a new state of the art among one-stage multi-person HMR methods and generalizes across benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of the method and to polish the text for readability. We did not use generative AI to generate synthetic datasets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in writing proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; assist with translation; clean or reformat datasets; support qualitative or thematic data analysis; or interpret results. We reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide hyperparameter configurations and implementation details of our experiments in Appendix C. We also include the source code in our submission, enabling reproduction of the results.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014. doi: 10.1109/CVPR.2014.471.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: Multi-person Whole-Body Human Mesh Recovery in a Single Shot. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 202–218, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73337-6. doi: 10.1007/978-3-031-73337-6_12.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.00843.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8. doi: 10.1007/978-3-030-58452-8_13.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2511.16719.

Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-Supervised Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2607–2616, 2022. doi: 10.1109/CVPR52688.2022.00264.

Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14586–14597, 2023. doi: 10.1109/ICCV51070.2023.01345.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is Secretly an Image Segmentation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25303–25313, 2025. doi: 10.1109/CVPR52734.2025.02356.

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A Video Dataset for In-The-Wild Close Human Interactions. In *Advances in Neural Information Processing Systems 37*, pp. 107270–107285, Vancouver, BC, Canada, 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS). ISBN 979-8-3313-1438-5. doi: 10.52202/079017-3407.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image Segmentation As Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9796–9805, 2020. doi: 10.1109/CVPR42600.2020.00982.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00371.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11107–11117, 2021. doi: 10.1109/ICCV48922.2021.01094.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 590–606, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20065-6. doi: 10.1007/978-3-031-20065-6_34.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2201.12329.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, October 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative Proxemics: A Prior for 3D Social Interaction from Images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9687–9697, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00925.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2304.07193.

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in Geography Optimized for Regression Analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13463–13473, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-6654-4509-2. doi: 10.1109/CVPR46437.2021.01326.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10967–10977, 2019. doi: 10.1109/CVPR.2019.01123.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2408.00714.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, December 2024. doi: 10.1145/3680528.3687565.

Jan Stenum, Melody M. Hsu, Alexander Y. Pantelyat, and Ryan T. Roemmich. Clinical gait analysis using video-based pose estimation: Multiple perspectives, clinical populations, and measuring change. *PLOS Digital Health*, 3(3):e0000467, March 2024. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000467.

Chi Su, Xiaoxuan Ma, Jiajun Su, and Yizhou Wang. SAT-HMR: Real-Time Multi-Person 3D Mesh Estimation via Scale-Adaptive Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16796–16806, 2025. doi: 10.1109/CVPR52734.2025.01565.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1834–1843, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.00180.

Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11159–11168, October 2021. doi: 10.1109/ICCV48922.2021.01099.

Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting People in their Place: Monocular Regression of 3D People in Depth. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13233–13242, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01289.

Mikael Svenstrup, Soren Tranberg, Hans Jorgen Andersen, and Thomas Bak. Pose estimation and adaptive robot behaviour for human-robot interaction. In *2009 IEEE International Conference on Robotics and Automation*, pp. 3571–3576, May 2009. doi: 10.1109/ROBOT.2009.5152690.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D Human Pose Estimation via Intuitive Physics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4713–4725, June 2023. doi: 10.1109/CVPR52729.2023.00457.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, volume 11214, pp. 614–631. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01249-6. doi: 10.1007/978-3-030-01249-6_37.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. PromptHMR: Promptable Human Mesh Recovery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1148–1159. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00115.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. Large-Scale Datasets for Going Deeper in Image Understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1480–1485, July 2019. doi: 10.1109/ICME.2019.00256.

Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. A Survey on Human Performance Capture and Animation. *Journal of Computer Science and Technology*, 32(3):536–554, May 2017. ISSN 1000-9000, 1860-4749. doi: 10.1007/s11390-017-1742-y.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. SAM 3D Body: Robust Full-Body Human Mesh Recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7209–7219, 2026. doi: 10.48550/arXiv.2602.15989.

Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D Instance Segmentation of Close Human Interaction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17016–17027, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01632.

Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M. Ni, and Lei Zhang. MP-Former: Mask-Piloted Transformer for Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18074–18083, 2023. doi: 10.1109/CVPR52729.2023.01733.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs Beat YOLOs on Real-time Object Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, Seattle, WA, USA, June 2024. IEEE. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01605.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5738–5746, 2019. doi: 10.1109/CVPR.2019.00589.

A DATASETS

Table 3: Training and evaluation datasets. "Downsample" is the subsampling stride applied to the datasets. "Filter" indicates whether the annotation-completeness filter described below is applied. "Available" denotes the number of images with 3D body annotations before subsampling and filtering, and "Used" denotes the number of images included in training or evaluation.

Dataset	Domain	3D labels	Downsample	Filter	Available	Used
TRAINING						
BEDLAM	Synthetic	Ground truth	2	N	286K	143K
AGORA	Synthetic	Ground truth	-	N	14K	14K
MOYO	Mocap	Ground truth	30	N	932K	29K
COCO	In-the-wild	SAM 3D Body	-	Y	24K	15K
MPII	In-the-wild	SAM 3D Body	-	Y	5K	3K
AIC	In-the-wild	SAM 3D Body	-	Y	170K	109K
SA-1B	In-the-wild	SAM 3D Body	6	Y	1,273K	106K
Total					2,704K	419K
EVALUATION						
Harmony4D	Interaction	Ground truth	20	-	288K	15K
Hi4D	Interaction	Ground truth	-	-	15K	15K
3DPW	In-the-wild	Ground truth	-	-	25K	25K
EMDB-1	In-the-wild	Ground truth	-	-	24K	24K
Total					352K	79K

Table 3 summarizes our training and evaluation datasets. BEDLAM (Black et al., 2023) and AGORA (Patel et al., 2021) are synthetic multi-person datasets with ground-truth SMPL annotations. MOYO (Tripathi et al., 2023) is a single-subject yoga dataset that provides examples of extreme poses and self-contact. BEDLAM and MOYO are released as videos, so we temporally subsample them to reduce temporal redundancy.

For the in-the-wild datasets COCO (Lin et al., 2014), MPII (Andriluka et al., 2014), AIC (Wu et al., 2019), and SA-1B (Kirillov et al., 2023), we use the SAM 3D Body (Yang et al., 2026) annotations. SAM 3D Body annotations provide meshes in the Momentum Human Rig body model, so we refit the meshes to gender-neutral SMPL. For some images in the dataset, the SAM 3D Body annotations cover only a subset of the people present. Because Mask-HMR is supervised to detect all human instances in the image, the human instances that are left unlabeled can act as false negatives and degrade recall during training. We thus run a completeness filter on the datasets, specifically, we run the RT-DETR (Zhao et al., 2024) person detector on the datasets and keep only images with no unannotated subjects. For SA-1B, we also filter for images with at least three annotated subjects to enable learning from crowded multi-person scenes.

3DPW (von Marcard et al., 2018) and EMDB (Kaufmann et al., 2023) are both standard in-the-wild datasets with 3D mesh annotations. Hi4D (Yin et al., 2023) and Harmony4D (Khirodkar et al., 2024) are datasets with close interactions between two people.

B CAMERA MODEL

We assume a full-perspective camera with a fixed field of view $\phi = 60^\circ$ along the longer side of the image (Baradel et al., 2025; Sun et al., 2022). Letting S denote the length of the longer image side, the focal length is defined as $f = \frac{S}{2\tan(\phi/2)}$. Directly regressing the metric root translation $\mathbf{t}_i = (t_{x,i}, t_{y,i}, t_{z,i})$ from query $\mathbf{q}_i$ is challenging. Under perspective projection, a larger person farther from the camera projects the same as a smaller person closer to it. Motivated by CLIFF (Li et al., 2022), each query instead predicts the normalized image-plane root coordinates and apparent scale of the corresponding person, $\mathbf{c}_i = (c_{x,i}, c_{y,i}, s_i)$, from which the root translation is recovered as

$$\mathbf{t}_i = \left(\frac{c_{x,i}}{s_i}, \frac{c_{y,i}}{s_i}, \frac{2f}{s_i\, S} \right). \tag{6}$$

Intuitively, s_i describes the scale at which a 2 m reference box would appear in the image at the depth of the person. The predicted image-plane coordinates $c_{x,i}, c_{y,i}$ and scale s_i therefore also define the region-of-interest bounding box of the associated human instance used by the implicit mask head (Section 3.2.3) and the per-joint pose refinement module (Section 3.2.4).

C TRAINING DETAILS

We resize each image while preserving its aspect ratio so that its longer side is $S = 896$ pixels. We train Mask-HMR for 60 epochs with AdamW, a weight decay of 10^{-4}, and a cosine learning-rate schedule. The initial learning rates are 2×10^{-5} for the pretrained ViT encoder and 4×10^{-5} for the remaining parameters.

Following previous DETR-based methods (Zhang et al., 2023; Su et al., 2025; Sun et al., 2024), we use deep supervision during training. The losses are applied to the final and intermediate decoder layers, with matching recomputed at each layer.

C.1 DETAILED TRAINING OBJECTIVES

C.1.1 BIPARTITE MATCHING

Following DETR (Carion et al., 2020), we assign predicted queries to ground-truth people using Hungarian matching. The matching cost combines query confidence, pointwise instance-mask BCE and Dice costs, and 2D-joint error, with weights of 2, 5, 5, and 5, respectively. For matching, the mask terms are evaluated at $K_{\text{match}} = 1024$ points.

C.1.2 TRAINING LOSSES

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{mesh}} + \mathcal{L}_{\text{mask}}. \tag{7}$$

The confidence loss is evaluated over all queries, with matched queries assigned positive targets and unmatched queries assigned negative targets. The mesh and mask losses are evaluated over matched queries. Unless stated otherwise, all loss terms are L_1 losses.

Confidence loss. The confidence objective is

$$\mathcal{L}_{\text{conf}} = \lambda_{\text{conf}} \mathcal{L}_{\text{focal}}, \tag{8}$$

where $\mathcal{L}_{\text{focal}}$ is the focal loss (Lin et al., 2017) with the class-balancing parameter $\alpha = 0.25$ and the focusing parameter $\gamma = 2$, and we set $\lambda_{\text{conf}} = 6$.

Mesh recovery losses. The mesh-recovery objective is

$$\mathcal{L}_{\text{mesh}} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{V}\mathcal{L}_{V} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}. \tag{9}$$

Here, $\mathcal{L}_{3D}$ and $\mathcal{L}_{2D}$ supervise the 3D joints and the 2D joint reprojections, $\mathcal{L}_{\text{pose}}$ and $\mathcal{L}_{\beta}$ supervise the SMPL pose and shape parameters, $\mathcal{L}_{V}$ supervises the SMPL vertices, $\mathcal{L}_{\text{depth}}$ supervises the root depth, and $\mathcal{L}_{\text{prior}}$ applies L_2 regularization for the predicted SMPL shape. We set $(\lambda_{3D}, \lambda_{2D}, \lambda_{\text{pose}}, \lambda_{\beta}, \lambda_{V}, \lambda_{\text{depth}}, \lambda_{\text{prior}}) = (10, 60, 5, 2, 1, 0.5, 0.1)$.

Mask losses. The predictions of the generated pointwise heads are supervised at the same K image points for all matched queries. The pointwise mask objective is

$$\mathcal{L}_{\text{mask}} = \underbrace{\lambda_{\text{bce}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}}_{\text{instance mask}} + \underbrace{\lambda_{\text{ce}}\mathcal{L}_{\text{CE}} + \lambda_{\text{pd}}\mathcal{L}_{\text{part-dice}}}_{\text{body-part masks}}. \tag{10}$$

For the instance mask, let $p_k \in [0, 1]$ denote the predicted mask probability and $y_k \in \{0, 1\}$ its binary target at sampled point k. The binary cross-entropy and Dice losses are $\mathcal{L}_{\text{BCE}}(p, y) = -\frac{1}{K}\sum_{k=1}^{K}[y_k \log p_k + (1 - y_k)\log(1 - p_k)]$ and $\mathcal{L}_{\text{Dice}}(p, y) = 1 - \frac{2\sum_k p_k y_k + 1}{\sum_k p_k + \sum_k y_k + 1}$, respectively. Cross-entropy losses measure pointwise classification error, whereas Dice losses measure the overlap between the predicted and ground-truth masks. For the body-part mask term, the cross-entropy and Dice losses are averaged over the N_P body-part channels. We set $(\lambda_{\text{bce}}, \lambda_{\text{dice}}, \lambda_{\text{ce}}, \lambda_{\text{pd}}) = (20, 5, 0.5, 0.5)$.

Table 4: Mask-loss point-budget (K) ablation on the BEDLAM and AGORA validation sets. We report NMVE (mm) and mask AP at an early training epoch (10) and the final training epoch (40). **Best** and second-best results in each column are **bolded** and underlined, respectively.

Points K	BEDLAM				AGORA			
	epoch 10		epoch 40		epoch 10		epoch 40	
	NMVE ↓	mask AP ↑	NMVE ↓	mask AP ↑	NMVE ↓	mask AP ↑	NMVE ↓	mask AP ↑
10	86.5	77.6	68.1	82.9	112.0	48.6	86.6	56.8
100	83.1	82.0	**65.8**	86.2	109.1	55.5	84.3	62.0
300	81.4	**84.0**	66.0	86.6	102.1	57.6	**83.7**	63.5
500	**81.1**	**84.0**	66.0	**86.8**	**100.5**	**58.3**	83.9	**63.9**

Table 5: Person-class instance segmentation on COCO. Scale: the person size relative to image area. **Best** and second-best results in each column are **bolded** and underlined, respectively.

Model	0–10%		10–20%		20–30%		30%+	
	mask-AP ↑	mIoU ↑	mask-AP ↑	mIoU ↑	mask-AP ↑	mIoU ↑	mask-AP ↑	mIoU ↑
SAM 3	**48.6**	77.5	70.0	85.1	72.1	86.6	74.6	87.3
EoMT-L	38.3	72.4	**74.8**	**86.4**	**78.0**	**87.8**	**82.8**	**89.7**
Mask-HMR	25.9	**81.4**	67.9	86.0	71.2	87.0	73.8	87.9

D Additional ablations

D.1 Effect of the number of sampled points

In this ablation, we report mask Average Precision (mask AP), which measures how accurately the model detects human instances and how well the predicted masks overlap with the ground truth.

In Table 4, we vary the number of points K sampled for the mask loss and report NMVE and mask AP over training. Models trained with more sampled points converge faster: at epoch 10, $K = 500$ improves NMVE over the $K = 10$ baseline by 5.4 mm on BEDLAM and 11.5 mm on AGORA, while the gap narrows to 2.1 mm and 2.7 mm by epoch 40. Pose accuracy saturates beyond $K = 100$, whereas mask AP keeps improving with more points, most visibly on the crowded AGORA scenes. We adopt $K = 300$ in all other experiments as a favorable accuracy-cost tradeoff.

E Additional qualitative results

SOTA comparisons. Figure 5 presents visual comparisons between our method and existing state-of-the-art approaches (Su et al., 2025; Baradel et al., 2025) on real-world images. Our method demonstrates strong generalization, producing accurate estimates across various scenarios, particularly under occlusion.

Failure cases. We also present failure cases of our model, as shown in Figure 6. The model produces inaccurate mesh estimates in challenging scenes with heavy occlusion and complex human poses. Figure 6 also shows that close interactions can lead to interpenetration, which results from the depth ambiguity inherent in monocular images.

F Segmentation results

Table 5 compares the segmentation performance of Mask-HMR with other state-of-the-art segmentation models EoMT (Kerssies et al., 2025) and SAM 3 (Carion et al., 2026) on the COCO validation set (Lin et al., 2014). We report both mask AP and matched mask IoU (mIoU) across different scales of the human instances. Compared with mask AP, mIoU measures the predicted mask quality without accounting for the detection accuracy. Mask-HMR is not able to detect very small people, which

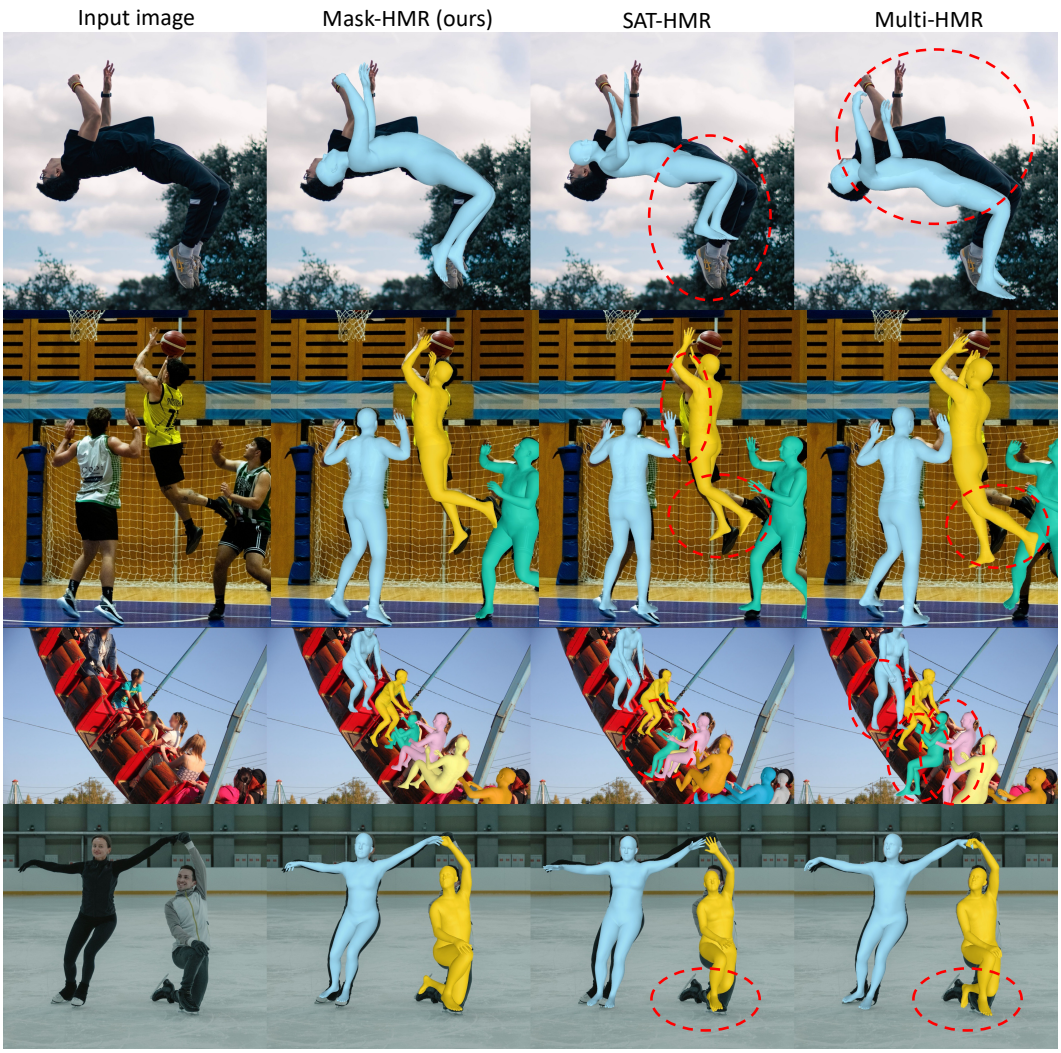

Figure 5: Qualitative comparison between models. Red dashed circles highlight instances of 2D misalignment or missed detections.

leads to a worse Mask-AP for human instances with a scale below 10%. This is because most small people are not annotated in the training images. However, for people at a regular scale, the model achieves segmentation performance comparable with that of other state-of-the-art dedicated segmentation methods. More qualitative results can be found in Figure 7.

G METRICS

Mean Per-Joint Position Error (MPJPE). MPJPE is the mean Euclidean distance between corresponding predicted and ground-truth 3D joints after pelvis alignment. It measures 3D pose accuracy.

Procrustes-Aligned MPJPE (PA-MPJPE). PA-MPJPE computes MPJPE after aligning each predicted person to the ground truth with a Procrustes Alignment. It measures local articulated-pose error independent of global translation, rotation, and scale.

Paired PA-MPJPE (Pair-PA-MPJPE). Pair-PA-MPJPE aligns both people jointly before computing MPJPE. In addition to per-person error, it also captures errors in the relative positions and orientations of the two persons.



Figure 6: Failure cases of our method. Mask-HMR can fail in cases involving extreme poses and heavy occlusion (top). The model also struggles with interpenetration in close interactions due to the depth ambiguity inherent in monocular images (bottom).

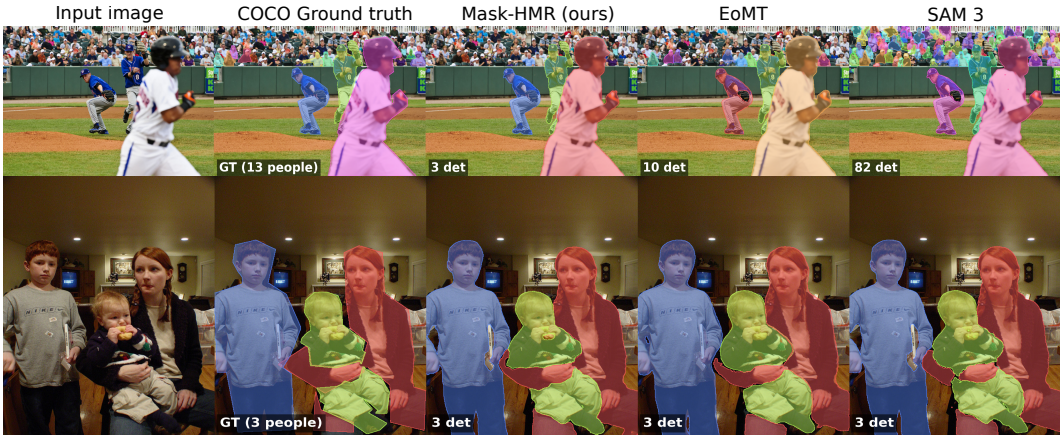

Figure 7: Qualitative comparison of segmentation performance on the COCO validation set. Det: number of human instances detected in the image.

Mean Vertex Error (MVE). MVE is the mean Euclidean distance between corresponding predicted and ground-truth SMPL vertices after pelvis alignment. Compared to MPJPE, it measures the combined pose and shape error.

Normalized Mean Vertex Error (NMVE). NMVE is MVE divided by the detection F1 score, penalizing missed detections and false-positive predictions. It jointly measures mesh reconstruction and human-instance detection accuracy.

Mask IoU (mIoU). The average overlap between a predicted mask and the annotated person it matches. Compared with mask AP, it measures mask accuracy on the human instances without accounting for detection performance.

Mask Average Precision (mask AP). Mask AP computes average precision over the precision–recall curve and averages across mask IoU thresholds from 0.50 to 0.95. It reflects both human-instance detection and mask-overlap quality.

All reconstruction errors are measured in millimeters (mm), with lower values indicating better performance. Higher mask AP and mIoU are better.