# RECONSTRUCTING HUMANS WITH ARTICULATED HANDS USING TRANSFORMERS

**Anonymous authors**
Paper under double-blind review

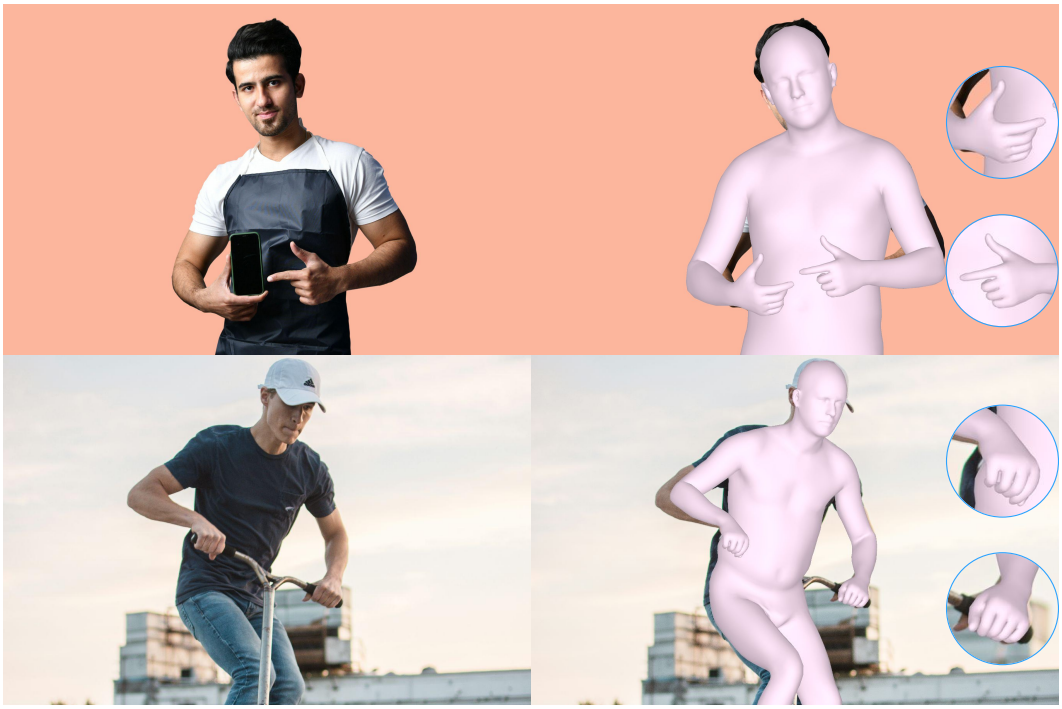

Figure 1: **Reconstructing expressive humans with articulated hands in 3D**. We propose BodhaHMR, an approach for **Bod**y and **Ha**nd **H**uman **M**esh **R**ecovery from a single input image. BodhaHMR delivers consistent estimates of body and hand pose by leveraging two dedicated backbones, one for body processing and one for hand processing. The features produced from each backbone are fused together by a unifying transformer decoder to predict the body and hand pose parameters in the form of the SMPL-H model (Romero et al., 2017).

## ABSTRACT

In this paper, we introduce an approach to reconstruct 3D humans with expressive hands given a single image as input. Current methods for pose estimation display robust performance for either bodies or hands. Unfortunately, these methods fail to simultaneously produce accurate 3D body and hand reconstructions. To address this limitation, we take a more cohesive approach to ensure both coarser and finer features of the human body are properly localized. Our approach is based on a feedforward network and following recent best practices, we adopt a fully transformer-based architecture. One of the key design choices we make is to leverage two separate backbone networks, one for 3D human pose and one for 3D hand pose estimation. These backbones process independently the body region and the hand regions and can make estimates about the bodies and the hands of the person. However, when the estimates are made independently, they tend to be inconsistent with one another and lead to unsatisfying reconstruction. Instead, we introduce a coupling transformer decoder that is trained to consolidate the intermediate features from the individual backbones into making a consistent estimate

for the body and the hands. The full system is trained on multiple datasets, including images with body ground truth, with hand ground truth, as well as images that include both body and hand ground truth. We evaluate our approach on the AGORA, ARCTIC, and COCO datasets, reporting metrics for both bodies and hands reconstruction accuracy to highlight our model's robustness over previous baselines.

## 1 INTRODUCTION

The reconstruction of expressive 3D human pose and shape from a monocular image requires a consistent understanding of various features of the presented human body, both coarse (e.g. head, torso, arms, legs) and fine (e.g. hands, fingers, feet, toes, eyes). Accordingly, a robust estimation method should appropriately respect features at both levels. When this is not the case, we observe clear discrepancies in the generated meshes. For instance, a system may implicitly prioritize fitting the predicted head and in turn sacrifice the prediction's hand pose and alignment, or vice versa. As human pose and shape estimation has numerous applications in robotics, healthcare, AR/VR, and sports, these inconsistencies can have noticeable consequences for downstream applications.

Breakthroughs in 3D pose estimation have recently relied on the robustness provided by coupling large models with large-scale data (Xu et al., 2022b; Goel et al., 2023; Pavlakos et al., 2024). This has also enabled significant progress in methods for 3D body (Goel et al., 2023; Dwivedi et al., 2024; Fiche et al., 2025; Su et al., 2025; Zhang et al., 2025) and 3D hand reconstruction (Pavlakos et al., 2024; Potamias et al., 2024; Chen et al., 2025; Fan et al., 2025). However, when it comes to jointly reconstructing humans with articulated hands, many recent approaches (Baradel* et al., 2024; Yin et al., 2025) tend to underperform. Specifically, they typically place less emphasis on finer features to better align coarser aspects of the body. Our aim is to address these limitations and make a more balanced and consistent 3D pose prediction by carefully considering coarse and finer features.

We present BodhaHMR, a method that takes crops of a person's body and hands from a single image as input, and performs **Bod**y and **Ha**nd **H**uman **M**esh **R**ecovery. In particular, our approach adopts two backbones: one for bodies and one for hands. These backbones take the corresponding crops as input and estimate features for the body and the two hands. The backbones are initialized using weights from a state-of-the-art approach for 3D body reconstruction (Goel et al., 2023) and a state-of-the-art approach for 3D hand reconstruction (Pavlakos et al., 2024). BodhaHMR then uses a coupling transformer decoder that takes as input the body and hand features and regresses the SMPL-H (Romero et al., 2017) parameters. As shown in Figure 1, BodhaHMR makes a unified and consistent estimate of a person's body and hands. Furthermore, BodhaHMR is powered by a large model and large data. BodhaHMR adopts a transformer-based architecture (Dosovitskiy et al., 2021; Xu et al., 2022b), while our training data consists of in-the-wild (Jin et al., 2020; Xu et al., 2022a), synthetic (Hewitt et al., 2024), and controlled (Zhu et al., 2023; Ionescu et al., 2014; 2011) datasets. 3D ground truth for bodies and hands is available in the synthetic and controlled data, while the in-the-wild data has 2D ground truth for hands and bodies. Together, these design decisions comprise a comprehensive approach for expressive mesh reconstruction.

We evaluate our method across a variety of datasets to demonstrate its robustness. For 3D pose (3D body and hand keypoints), we benchmark our approach on the ARCTIC (Fan et al., 2023) and AGORA (Patel et al., 2021) datasets. ARCTIC provides many challenging and expressive hand poses in a studio setting, while AGORA introduces synthetic subjects and environments, with some being low-resolution. For 2D pose, in addition to ARCTIC and AGORA, we benchmark on images from COCO-Wholebody (Jin et al., 2020; Xu et al., 2022a). COCO offers many diverse in-the-wild examples for our testing.

We contribute BodhaHMR, a cohesive and holistic approach for expressive human mesh recovery from a single RGB image. Moreover, we perform extensive evaluation on BodhaHMR and alternative state-of-the-art approaches. Our testing reveals that BodhaHMR outperforms state of the art across the board on 2D hand pose while maintaining state-of-the-art performance on bodies.

## 2 RELATED WORK

**3D human body pose and shape estimation.** Human mesh recovery has been considered from various perspectives, but we will limit our scope to approaches that derive pose and shape parameters from a single image. Previous optimization-based approaches estimate 3D human pose and shape by analyzing 2D image features (Guan et al., 2009; Bogo et al., 2016; Lassner et al., 2017; Omran et al., 2018; Pavlakos et al., 2019a; Xu et al., 2020). Since the introduction of HMR (Kanazawa et al., 2018), which learned to estimate SMPL parameters using a convolutional neural network (CNN), regression-based approaches have become more prevalent (Pavlakos et al., 2019b; Guler & Kokkinos, 2019; Jiang et al., 2020; Georgakis et al., 2020; Kolotouros et al., 2021; Kocabas et al., 2021; Zhang et al., 2021).

Recently, there has been a shift from CNNs to transformer-based (Vaswani et al., 2017) architectures. Mesh Graphformer (Lin et al., 2021) directly estimates the mesh vertices using transformers. HMR 2.0 (Goel et al., 2023), on the contrary, regresses SMPL parameters without domain-specific design decisions to outperform previous baselines. TokenHMR (Dwivedi et al., 2024), VQ-HPS (Fiche et al., 2024), and MEGA (Fiche et al., 2025) use a Vector Quantized-Variational Autoencoder (VQ-VAE) to tokenize human pose related information and minimize irregular predictions. ADHMR (Shen et al., 2025) iterates on existing diffusion-based HMR methods by performing direct preference optimization (Wallace et al., 2023). DeforHMR (Heo et al., 2025) uses a novel query-agnostic implementation of deformable attention transformers (Xia et al., 2023; 2022; Zhu et al., 2020) to enhance the model's spatial awareness. PAMA (Chen & He, 2025) couples a module for limb appearance consistency with full-perspective projection and an adapted reprojection loss to handle alignment discrepancies. Other approaches focus on optimizing for different challenges of human pose estimation. For instance, MetricHMR (Zhang et al., 2025) regresses 3D position information in addition to human pose and shape parameters, using a camera ray representation method. SAT-HMR (Su et al., 2025) tackles the issue of high computational costs in one-stage multi-person pose estimation. Our approach is different to these approaches by aiming to reconstruct not only the body, but also the hands of the person.

**3D hand pose and shape estimation.** Similarly to the above, we will focus on approaches that recover hand pose and shape parameters from a single image. Most of the initial methods (Baek et al., 2019; Zhang et al., 2019; Boukhayma et al., 2019; Park et al., 2022; Oh et al., 2023) for this problem employ the MANO parametric model (Romero et al., 2017) and regress the parameters. Other early approaches estimate the mesh vertices instead of regressing a parametric model (Ge et al., 2019; Chen et al., 2022; Jiang et al., 2023). As with body pose estimation, a similar shift to transformer architectures for hand pose estimation has occurred. HaMeR (Pavlakos et al., 2024) adopts a vanilla transformer-based regression network, surpassing previous approaches without a specialized design. Many of these transformer-based approaches make intentional augmentations to resolve critical challenges. WiLoR (Potamias et al., 2024) deploys a coarse-to-fine method by passing predicted parameters through a refinement module, to improve alignment and handle occlusions. To achieve high computational efficiency, simpleHand (Zhou et al., 2024) decomposes its mesh decoder into token generator and mesh regressor modules, and develops a streamlined structure. HHMR (Li et al., 2024b) uses a graph diffusion model in combination with attention mechanisms to construct a flexible and robust framework capable of many mesh recovery tasks. HandOS (Chen et al., 2025) builds a one-stage pipeline to detect hands and estimate pose, without directly regressing the MANO parameters. Fan et al. (2025) utilizes temporal information to improve performance on low-resolution images. To increase estimation accuracy, EHPE (Zheng et al., 2025) focuses on analyzing the distal phalanx tip and wrist joints, while Karvounas et al. (2025) incorporate texture-based supervision into existing estimation methods. Instead of transformers, Hamba (Dong et al., 2024) opts for a Mamba-based (Gu & Dao, 2023; Dao & Gu, 2024) architecture complete with graph learning and state space modeling. Although we use insights from the hand pose estimation literature and we adopt the HaMeR backbone (Pavlakos et al., 2024) in our architecture, we are different to these approaches by estimating a holistic body and hand reconstruction.

**3D expressive human mesh recovery.** The body approaches detailed so far focus on the SMPL parametric model, which does not capture the hand articulation or facial expressions. Earlier attempts at expressive mesh recovery (Choutas et al., 2020; Moon et al., 2022; Zhang et al., 2023) regress the SMPL-X parametric model (Pavlakos et al., 2019a) parameters to capture more nuances

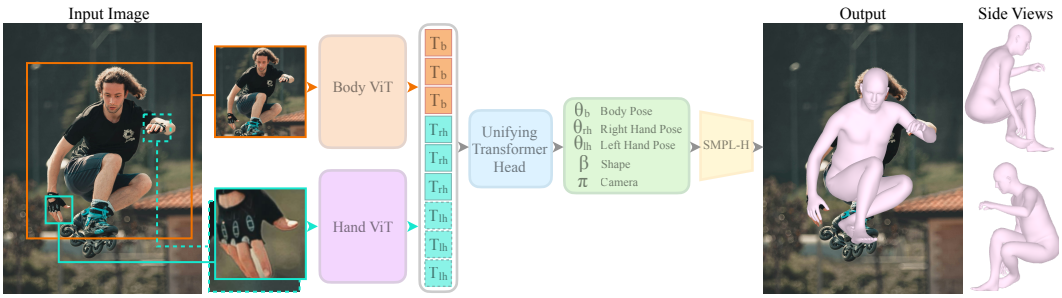


Figure 2: **Overview of BodhaHMR's architecture**. We present a transformer-based method of expressive human mesh recovery. Our network takes crops from an image of a person's body, right hand, and left hand as input, and passes these through two vision transformer backbones. Specifically, the body crop is passed through a dedicated backbone for bodies, while the hand crops are passed through a dedicated backbone for hands. The tokens generated from the body $T_b$, right hand $T_{rh}$, and the left hand $T_{lh}$ are concatenated along the token dimension and passed to a unifying transformer head, which ultimately outputs accurate SMPL-H (Romero et al., 2017) parameters.

in subjects' hands and faces. Recently, we have observed the same transformer shift in this context when it comes to the network architecture. SMPLer-X (Cai et al., 2023) and its follow up, SMPLest-X (Yin et al., 2025), primarily demonstrate the positive impacts of scaling up data and using large Vision Transformers (ViT) (Dosovitskiy et al., 2021; Xu et al., 2022b). OSX (Lin et al., 2023) designs a Component Aware Transformer (CAT) to provide a one-stage framework to retrieve face, hand, and body parameters that ensures good connectivity between subject attributes. Multi-HMR (Baradel* et al., 2024) and AiOS (Sun et al., 2024) present single-shot approaches to deterministically deal with the challenge of multi-person expressive pose estimation while retaining accurate 3D locations. On the other hand, CondiMen (Romain et al., 2025) deploys a probabilistic method to address multi-person pose estimation. PromptHMR (Wang et al., 2025) enables pose and shape estimation through both spatial and semantic prompts to preserve visual context. Moving away from transformers, D-PoSE (Vasilikopoulos et al., 2024) uses a lighter CNN backbone in combination with depth and part-segmentation maps to outperform competing transformer-reliant methods.

Following the trend in body, hand, and expressive pose estimation, we use a transformer architecture to recover body and articulated hand parameters. Unlike the vast majority of techniques for expressive mesh recovery, we choose to regress the SMPL-H parametric model (Romero et al., 2017) instead of SMPL-X (Pavlakos et al., 2019a). At a first glance, this looks like a more constrained setting, but we are motivated by the need for very precise body and hand pose estimation for multiple downstream applications, particularly in robotics (Li et al., 2024a; Fu et al., 2024; Wang et al., 2024).

## 3 Technical Approach

In this section, we present the technical details of our approach. First we provide some preliminaries about the SMPL-H model (Section 3.1), and then we elaborate on our approach (Section 3.2), the architecture we use (Section 3.3) and the training losses (Section 3.4).

### 3.1 SMPL-H model

For our reconstruction, we employ the SMPL-H parametric model (Romero et al., 2017) of the human body and hands. SMPL-H takes as input pose ($\theta \in \mathbb{R}^{52\times3\times3}$) and shape ($\beta \in \mathbb{R}^{10}$) parameters and outputs a mesh $M \in \mathbb{R}^{3\times V}$ of the human body with articulated hands, where $V = 6890$ vertices. The pose parameters include body pose parameters $\theta_b \in \mathbb{R}^{21\times3\times3}$, global orientation $\theta_g \in \mathbb{R}^{3\times3}$, and hand pose parameters for the left and the right hand ($\theta_{lh}, \theta_{rh} \in \mathbb{R}^{15\times3\times3}$). The joints $X \in \mathbb{R}^{3\times k}$ are a linear combination of the vertices.

### 3.2 HUMAN MESH RECONSTRUCTION WITH EXPRESSIVE HANDS

Our aim is to use an RGB image of a person to generate a representative, articulated 3D mesh. To accomplish this, we learn a predictor function $f(I)$ that maps the image to the SMPL-H parameters $\Theta = (\theta, \beta, \pi)$. This function also produces camera parameters $\pi = (R, t)$ as part of the output parameters. Our camera model has fixed focal length and camera intrinsics $K$. With global orientation $R \in \mathbb{R}^{3\times3}$ and translation $t \in \mathbb{R}^3$, we can project the 3D joints $X$ onto the image using $x = \pi(X) = \Pi(K(RX + t))$. Note that $\theta$ contains a global orientation, so we fix $R$ to be the identity in practice. Ultimately, we learn $f(I) = \Theta$.

### 3.3 ARCHITECTURE

Following Goel et al. (2023) and Pavlakos et al. (2024), our approach utilizes a "transformerized" design. As shown in Figure 2, the architecture comprises two Vision Transformer huge (ViT-H) backbones (Dosovitskiy et al., 2021) (one for bodies, one for hands) and a unifying transformer decoder head. The coupling decoder cohesively considers body and hand features before making a prediction. This significantly contrasts having separate decoders for the body and hands, which make predictions for the bodies and hands in isolation, and require us to fuse the estimates together afterwards. We take three crops of the bounding boxes of the body ($I_b$), right and left hands ($I_{rh}, I_{lh}$) from the original image and transform each one into $16 \times 16$ patches to feed into the corresponding backbones. The ViT-H backbones convert the patches into a total of three sets of tokens, one set per crop. We concatenate these token outputs along the token dimension and feed them all into the decoder, which cross-attends to them. Finally, the unifying transformer head outputs $\Theta$ corresponding to the given subject.

### 3.4 LOSSES

Our model is trained with a combination of 2D and 3D losses, mirroring best practices from prior literature (Kanazawa et al., 2018; Kolotouros et al., 2019; Goel et al., 2023). The datasets we train with contain a variety of annotations, so we deploy a subset of the losses below for each training example. If ground-truth SMPL-H pose ($\theta^*$) and shape ($\beta^*$) parameters are available, we use the following MSE loss:

$$\mathcal{L}_{smplh} = ||\theta_b - \theta_b^*||_2^2 + ||\theta_g - \theta_g^*||_2^2 + ||\theta_{rh} - \theta_{rh}^*||_2^2 + ||\theta_{lh} - \theta_{lh}^*||_2^2 + ||\beta - \beta^*||_2^2. \quad (1)$$

With this loss formulation, we can straightforwardly handle partial annotations by "turning off" missing terms. If ground-truth 3D joint annotations $X^*$ are available, we use the following L1 loss on the 3D keypoints:

$$\mathcal{L}_{kp3D} = ||X - X^*||_1. \quad (2)$$

If ground-truth 2D keypoint annotations $x^*$ are available, we use the following reprojection L1 loss on the projected 3D keypoints:

$$\mathcal{L}_{kp2D} = ||x - x^*||_1. \quad (3)$$

Lastly, it is possible that the model predicts abnormal 3D poses when ground-truth pose or joint annotations are not available. We combat this by training discriminators $D_k$ for the (i) body pose parameters $\theta_b$, (ii) shape parameters $\beta$, and (iii) each individual body joint angle, similar to Kanazawa et al. (2018). The corresponding generator loss is:

$$\mathcal{L}_{adv} = \sum_k (D_k(\theta_b, \beta) - 1)^2. \quad (4)$$

## 4 EXPERIMENTS

### 4.1 PRELIMINARIES

**Datasets and implementation.** We train our BodhaHMR on COCO-Wholebody (Jin et al., 2020; Xu et al., 2022a), Human3.6M 3D WholeBody (Zhu et al., 2023; Ionescu et al., 2014; 2011), and SynthMoCap (SynthBody and SynthHand) (Hewitt et al., 2024). These datasets offer a blend of in-the-wild, controlled, and synthetic examples to generate a robust final model. We utilize the pretrained weights of HMR 2.0b (Goel et al., 2023) and HaMeR (Pavlakos et al., 2024) for body

| Method | AGORA | | | | ARCTIC | | | | COCO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | @0.05↑ | @0.1↑ |
| Frankenstein (Hu, 2025) | 58.4 | 11.8 | 0.050 | 0.171 | 25.8 | 10.1 | 0.112 | 0.351 | 0.069 | 0.206 |
| SMPLest-X (Yin et al., 2025) | 51.0 | 10.0 | 0.012 | 0.047 | 41.9 | 15.7 | 0.013 | 0.052 | 0.012 | 0.047 |
| Multi-HMR 896L* (Baradel* et al., 2024) | - | - | - | - | 35.1 | 13.6 | 0.038 | 0.137 | 0.013 | 0.049 |
| Multi-HMR 672L* (Baradel* et al., 2024) | - | - | - | - | 40.3 | 13.0 | 0.036 | 0.126 | 0.011 | 0.042 |
| Ours | 47.2 | 10.8 | 0.105 | 0.312 | 30.2 | 11.3 | 0.121 | 0.365 | 0.120 | 0.361 |

Table 1: **Comparison with the state-of-the-art on hand reconstruction.** We evaluate the predicted 3D and reprojected 2D hand joints by taking the average of metrics from both hands. We report MPJPE, PA-MPJPE, PCK @0.05, and PCK @0.1 if 3D hand ground truth is available (AGORA and ARCTIC). If only 2D hand ground truth annotations are available (COCO), PCK @0.05 and PCK @0.1. Our method outperforms most state of the art in the 3D evaluation. The Frankenstein approach produces more accurate 3D hands in some cases because it copies hands from HaMeR (Pavlakos et al., 2024). BodhaHMR achieves the best 2D results across all datasets. * denotes methods trained on the AGORA training split. All MPJPE and PA-MPJPE metrics are in mm.

| Method | AGORA | | | | ARCTIC | | | | COCO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | @0.05↑ | @0.1↑ |
| Frankenstein (Hu, 2025) | 157.9 | 67.0 | 0.801 | 0.915 | 163.5 | 77.3 | 0.762 | 0.904 | 0.806 | 0.941 |
| SMPLest-X (Yin et al., 2025) | 101.4 | 61.1 | 0.657 | 0.855 | 95.6 | 41.2 | 0.761 | 0.948 | 0.548 | 0.831 |
| Multi-HMR 896L* (Baradel* et al., 2024) | - | - | - | - | 125.0 | 56.6 | 0.679 | 0.856 | 0.530 | 0.772 |
| Multi-HMR 672L* (Baradel* et al., 2024) | - | - | - | - | 115.9 | 53.4 | 0.679 | 0.881 | 0.473 | 0.713 |
| Ours | 142.0 | 61.6 | 0.765 | 0.909 | 108.3 | 51.5 | 0.755 | 0.949 | 0.718 | 0.915 |

Table 2: **Comparison with the state-of-the-art on body reconstruction.** We report results on body reconstruction accuracy. Our method maintains consistent state-of-the-art performance on 3D and 2D evaluation. * indicates methods trained on the AGORA training split. All MPJPE and PA-MPJPE numbers are in mm.

and hands backbones, respectively. These backbones provide good initialization for their respective tasks, as they have been trained on numerous body and hand examples. During training, we freeze these backbones and only train the unifying transformer head parameters. To encourage better alignment of the mesh hands to the image, we take inspiration from Pavlakos et al. (2019a) and Choutas et al. (2020), and gradually increase the 2D hand loss weights.

**Baselines.** For our evaluation, we compare against state-of-the-art methods for expressive mesh recovery. We examine the performance of two ViT-L (Dosovitskiy et al., 2021; Xu et al., 2022b) based approaches from Multi-HMR (Baradel* et al., 2024), which process input images at resolutions of 896x896 (896L) and 672x672 (672L), respectively. Furthermore, we report results from the publicly available checkpoint of SMPLest-X (Yin et al., 2025). Lastly, to compare with a straightforward approach of fusing the results from an independent network for bodies and an independent network for hands, we design another "Frankenstein" baseline. For each example, we use HMR 2.0 to estimate $\theta_b, \theta_g, \beta$ and HaMeR to estimate $\theta_{rh}, \theta_{lh}$. Straightforwardly generating a SMPL-H mesh with these parameters results in poor hand alignments due to discrepancies between the estimated hand wrist joints and the estimated body wrist joints. To achieve more consistent body and hand integration, we recompute the wrist poses by taking the elbow rotations from HMR 2.0, the hand rotations from the HaMeR, and applying the logic from Hu (2025).

**Metrics.** We employ the AGORA (Patel et al., 2021) and ARCTIC (Fan et al., 2023) allocentric validation sets to evaluate 3D pose prediction. In particular, we report Mean Per Joint Position Error before (MPJPE) (Ionescu et al., 2014) and after Procrustes Alignment (PA-MPJPE) (Kanazawa et al., 2018; Zhou et al., 2019). To evaluate 2D pose accuracy, we use the ARCTIC allocentric, AGORA, and COCO-Wholebody validation sets. We report the Percentage of Correct Keypoints (PCK) (Yang & Ramanan, 2013) metric at thresholds of 0.05 and 0.1 for the body and hands.

### 4.2 HAND RESULTS

We report the results of hand reconstruction in Table 1. In our evaluation, BodhaHMR model excels in the context of hand reconstruction. Our approach outperforms the other baselines on all given metrics, except for the 3D hand pose evaluation on ARCTIC and PA-MPJPE on AGORA. In these cases, we see that BodhaHMR does better than its competitors on the 2D pose evaluation. As the

| Models | AGORA | | | | ARCTIC | | | | COCO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | @0.05↑ | @0.1↑ |
| Single Backbone | 53.7 | 11.2 | 0.0283 | 0.106 | 47.5 | 17.7 | 0.0426 | 0.158 | 0.0416 | 0.156 |
| BodhaHMR | **51.1** | **10.3** | **0.0589** | **0.201** | **40.3** | **12.3** | **0.0528** | **0.185** | **0.0613** | **0.214** |

Table 3: **Ablation on hand backbone.** We study the impact of introducing the hand backbone into the network. The single backbone approach freezes the pretrained weights from Goel et al. (2023) during training and regresses the SMPL-H parameters from one crop of the subject. We compare this against BodhaHMR on hand reconstruction. The single backbone version performs consistently worse than our method on these benchmarks, highlighting the importance of the hand backbone.

2D pose evaluation measures the reprojection of the 3D joints on to the image, BodhaHMR's hand localization on the image is more accurate than Frankenstein and SMPLest-X, respectively. This demonstrates the effectiveness of BodhaHMR's consolidated approach: our method consistently estimates more accurate 2D hand positioning than competitors. Moreover, for the 2D hand metrics, we observe that Frankenstein is consistently the second-best approach. Our method realizes gains of 1.5× to 2× over Frankenstein for AGORA and COCO. On ARCTIC, BodhaHMR outperforms Frankenstein by 3% to 7%. The in-the-wild images from the COCO dataset have more subjects, interactions, environments, and perspectives than the studio images from ARCTIC. AGORA also presents images with varied and numerous subjects in a variety of settings. In contrast, ARCTIC's validation split only contains one subject in a controlled, studio setting. Therefore, we believe 2D performance on COCO and AGORA provides a better indication of a model's generalizability and robustness than ARCTIC.

### 4.3 Body results

We report the results of body reconstruction in Table 2. As we see, our approach lags behind only SMPLest-X, and beats out the other methods in 3D body pose. Note that the body joints of the SMPL-H and SMPL-X parametric models are localized differently, and AGORA and ARCTIC provide ground-truth annotations with respect to the SMPL-X model. Despite this discrepancy, BodhaHMR's performance on the 3D body evaluation competes with state-of-the-art methods that regress SMPL-X parameters, demonstrating the robustness of our model. Similarly, on AGORA and COCO 2D evaluation, BodhaHMR outperforms SMPLest-X and Multi-HMR and slightly falls behind the Frankenstein method. Given these results, BodhaHMR maintains state-of-the-art performance on the body metrics. This is significant when contextualized with the hand results above: BodhaHMR produces the most consistent and balanced predictions across all tested methods.

### 4.4 Ablation study

We investigate the impact of utilizing distinct backbones for hands and bodies on the hand accuracy in Table 3. To examine this, we compare against a model that adopts the encoder backbone from Goel et al. (2023) and freezes these weights while training. At test time, it passes a crop of a person's body through the backbone, which produces one set of image tokens as output. These tokens are passed through a standard decoder head, which cross-attends to them and regresses the SMPL-H parameters. As we see, our method outperforms the straightforward extension of HMR 2.0 across the board: BodhaHMR provides more accurate 3D hand pose reconstructions and better 2D alignments. Our method is superior in estimating both the hand pose and 2D hand location than the simpler model, justifying the use second backbone for the hands. Please refer to Section A.4 for more implementation details and body results of this ablation.

### 4.5 Qualitative results

We perform a qualitative comparison of our approach with state-of-the-art in Figure 3. These results support our findings from the quantitative evaluation, as BodhaHMR demonstrates more accurate 2D hand location in combination with similar body estimates as the other methods. In fact, due to BodhaHMR's unified approach, it achieves significantly better image alignment in the egocentric example. In Figures 4 and 5, we provide more qualitative results from BodhaHMR. Our method produces well-aligned bodies and hands under various viewpoints, occlusions, and environments.

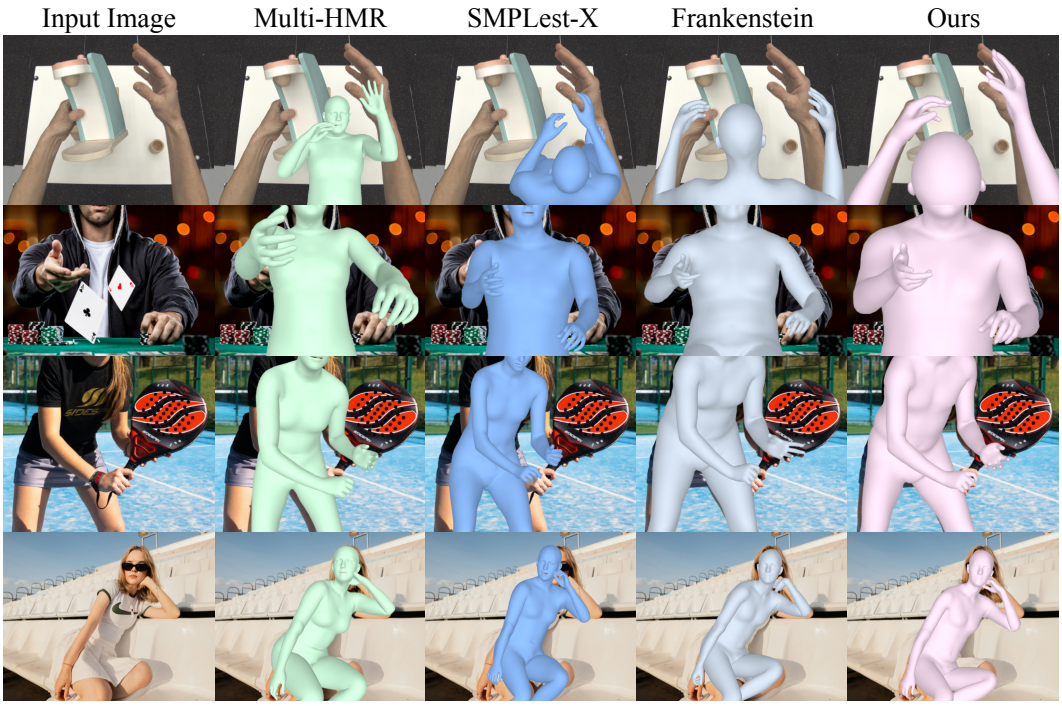


Figure 3: **Qualitative comparison with state-of-the-art**. We show results from the 896L version of Multi-HMR. In challenging egocentric and in-the-wild settings, our model recovers more accurate hand alignments and articulations than competitors. Please zoom in for details.

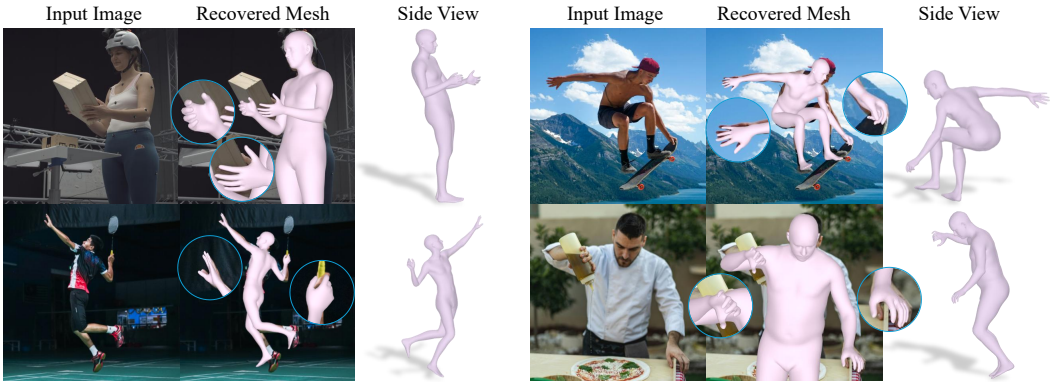


Figure 4: **Novel views of BodhaHMR**. For each input image, we show the recovered mesh, hand close-ups, and a side view. Our method gracefully handles a variety of poses and occlusions.

## 5 Conclusion

We describe BodhaHMR, a method for reconstructing 3D expressive humans from a single RGB image. Our approach builds on state-of-the-art methods for bodies and hands: BodhaHMR consolidates the body and hand features from these networks to achieve notable improvements on 2D hand alignment, without sacrificing state-of-the-art performance on the body.

**Limitations.** One limitation of BodhaHMR is we notice some of the reconstructed hands could be more expressive and located even more precisely. This may be the result of the unifying decoder incorrectly localizing the tokens from the hand backbones on the input image. Future work could provide more context to the unifying transformer head using positional embeddings (Prakash et al., 2024) to potentially mitigate these issues. Additionally, while we provide some qualitative egocen-


Figure 5: **Qualitative results of BodhaHMR**. For each input image, we display the overlaid reconstruction, close-ups of the input hands, and their reconstructions. Rows 1-2 are from ARCTIC, row 3 is from AGORA, and rows 4-7 are from the Internet. Our approach recovers cohesive and expressive hands alongside robust bodies.

tric examples with positive results, our method could perform even more robustly in this context. We observe good performance due to the synthetic hand training data, which contains isolated hand examples that effectively mimic egocentric images. Future work could explore explicitly adding more egocentric examples and datasets to improve performance in this context.

## REFERENCES


Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1067–1076, 2019. doi: 10.1109/CVPR.2019.00116.

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 *ECCV*, 2024.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer International Publishing, October 2016.

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10843–10852, 2019.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023.

Shu Chen and Ying He. Knowledge-embedded transformer for 3-d human pose estimation. *IEEE Transactions on Instrumentation and Measurement*, 74:1–11, 2025. doi: 10.1109/TIM.2025.3569914.

Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20512–20522, 2022. doi: 10.1109/CVPR52688.2022.01989.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. Handos: 3d hand reconstruction in one stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, pp. 20–40, 2020. URL `https://expose.is.tue.mpg.de`.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando D De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *Advances in Neural Information Processing Systems*, 37:2127–2160, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kaixin Fan, Pengfei Ren, Jingyu Wang, Haifeng Sun, Qi Qi, Zirui Zhuang, and Jianxin Liao. Pose-guided temporal enhancement for robust low-resolution hand reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22627–22637, 2025. doi: 10.1109/CVPR52734.2025.02107.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. VQ-HPS: Human pose and shape estimation in a vector-quantized latent space. In *European Conference on Computer Vision (ECCV)*, 2024.

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10825–10834, 2019. doi: 10.1109/CVPR.2019.01109.

Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision*, pp. 768–784. Springer, 2020.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Peng Guan, Alexander Weiss, Alexandru O. Bălan, and Michael J. Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1381–1388, 2009. doi: 10.1109/ICCV.2009.5459300.

Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10884–10894, 2019.

Jaewoo Heo, George Hu, Zeyu Wang, and Serena Yeung-Levy. DeforHMR: Vision Transformer with Deformable Cross-Attention for 3D Human Mesh Recovery . In *2025 International Conference on 3D Vision (3DV)*, pp. 1594–1604, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/3DV66043.2025.00149. URL https://doi.ieeecomputersociety.org/10.1109/3DV66043.2025.00149.

Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024.

Wentao Hu. Mano2smpl-x. https://github.com/VincentHu19/Mano2Smpl-X, 2025.

Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *International Conference on Computer Vision*, 2011.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2020.

Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 758–767, 2023.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

Giorgos Karvounas, Nikolaos Kyriazis, Iason Oikonomidis, Georgios Pavlakos, and Antonis A. Argyros. Enhancing monocular 3d hand reconstruction with learned texture priors, 2025. URL https://arxiv.org/abs/2508.09629.

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

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 11127–11137, October 2021.

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

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11605–11614, 2021.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL http://up.is.tuebingen.mpg.de.

Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024a.

Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 645–654, 2024b. doi: 10.1109/CVPR52733.2024.00068.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022.

Yeounguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, Verona, Italy, 2018.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019a.

Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 803–812, 2019b.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024.

Aditya Prakash, Arjun Gupta, and Saurabh Gupta. Mitigating perspective distortion-induced shape ambiguity in image crops. In *European Conference on Computer Vision (ECCV)*, 2024.

Brégier Romain, Baradel Fabien, Lucas Thomas, Galaaoui Salma, Armando Matthieu, Weinzaepfel Philippe, and Rogez Grégory. Condimen: Conditional multi-person mesh recovery, 2025. URL https://arxiv.org/abs/2412.13058.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):1–17, November 2017. ISSN 1557-7368. doi: 10.1145/3130800.3130883. URL http://dx.doi.org/10.1145/3130800.3130883.

Wenhao Shen, Wanqi Yin, Xiaofeng Yang, Cheng Chen, Chaoyue Song, Zhongang Cai, Lei Yang, Hao Wang, and Guosheng Lin. Adhmr: Aligning diffusion-based human mesh recovery via direct preference optimization. In *International Conference on Machine Learning*. PMLR, 2025.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 16796–16806, June 2025.

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, 2024. doi: 10.1109/CVPR52733.2024.00180.

Nikolaos Vasilikopoulos, Drosakis Drosakis, and Antonis Argyros. D-pose: Depth as an intermediate representation for 3d human pose and shape estimation. *arXiv preprint arXiv:2410.04889*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4803, June 2022.

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6183–6192, 2020. doi: 10.1109/CVPR42600.2020.00622.

Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: Searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022b.

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013. doi: 10.1109/TPAMI.2012.261.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *arXiv preprint arXiv:2501.09782*, 2025.

He Zhang, Chentao Song, Hongwen Zhang, and Tao Yu. Metrichmr: Metric human mesh recovery from monocular images, 2025. URL https://arxiv.org/abs/2506.09919.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2354–2364, 2019. doi: 10.1109/ICCV.2019.00244.

Bolun Zheng, Xinjie Liu, Qianyu Zhang, Canjin Wang, Fangni Chen, and Mingen Xu. Ehpe: A segmented architecture for enhanced hand pose estimation, 2025. URL https://arxiv.org/abs/2507.09560.

Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):901–914, April 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2816031. URL https://doi.org/10.1109/TPAMI.2018.2816031.

Zhishan Zhou, Shihao. zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20166–20177, October 2023.

# A Appendix

## A.1 Architecture

As detailed in the main manuscript, BodhaHMR uses two ViT-H (Dosovitskiy et al., 2021) image encoders. We take the backbone for bodies from HMR 2.0 (Goel et al., 2023) and the backbone for hands from HaMeR (Pavlakos et al., 2024). Both backbones accept input images of size $256 \times 192$ and output $16 \times 12$ image tokens, each of dimension 1280. At test time, we pass a crop of the person's body to the body backbone and the crops of their hands sequentially to the hands backbone. The three sets of tokens produced by the encoders are concatenated along the token dimension and passed to the coupling transformer decoder. For the unifying decoder, we adopt the same architecture as Goel et al. (2023). It has 6 layers and takes a 1024-dimensional SMPL-H query token as input and cross-attends to the concatenated image tokens. The unifying decoder has 8 (64- dim) heads for cross-attention and self-attention and a hidden dimension of 1024. We readout $\theta_b, \theta_g, \theta_{rh}, \theta_{lh}\, \beta$, and $\pi$ from the output token of the coupling decoder.

## A.2 Training

During training, we use a weight of 0.13 to sample examples from Human3.6M, 0.12 for Human3.6M with hands, 0.25 for SynthBody, 0.25 for SynthHand, 0.13 for COCO, and 0.12 for COCO with hands. Note that we have two versions of the COCO and Human3.6M datasets: one version containing only body annotations and one version with body and hand annotations. Moreover, we train our final model with AdamW (Kingma & Ba, 2014) on four Nvidia A6000 GPUs with an effective batch size of $512 \times 4 = 2048$ for almost 500k iterations. We utilize a learning rate of $1\mathrm{e}{-5}$, a weight decay of $1\mathrm{e}{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Following best practices (Goel et al., 2023), we perform augmentations while training. These include randomly rescaling the size and color, flipping, rotating, and translating the center of the body bounding box. To avoid passing in hand crops of extremely low resolutions, we do not rescale the size of the hand bounding boxes. We apply the remaining augmentations on the hands. Across all iterations, we use a weight of 0.05 for 3D keypoint loss, 0.001 on the loss for $\theta_b, \theta_g, \theta_{rh}, \theta_{lh}$, $5\mathrm{e}{-4}$ on the loss for $\beta$, $5\mathrm{e}{-4}$ for adversarial loss and 0.01 for 2D body keypoint loss. As mentioned in the main text, we gradually increase the loss weights on the 2D hand keypoints. We set the loss weight of the 2D hand keypoints at 0.01 for the first 190k iterations of training. Then, for the next 294k iterations, we boost it to 0.2 to encourage better 2D hand alignment. To process left hands, we adopt the same left-right flipping from Pavlakos et al. (2024).

## A.3 Evaluation

The Multi-HMR (Baradel* et al., 2024) methods are designed to detect and estimate multiple expressive humans in one-stage. In contrast, all other tested approaches accept crops of each person in an image and individually predict the corresponding human pose parameters. To evaluate the Multi-HMR approaches in the same context as the other methods, we force predictions of individuals. Multi-HMR requires the person's 2D head keypoint to force a predictions on a particular person. In our testing, we estimate this by taking the average of the ground truth ear keypoints. Since AGORA (Patel et al., 2021) and ARCTIC (Fan et al., 2023) provide ground truth ear keypoints for all subjects, we report metrics on the entirety of those validation sets. However, the COCO-Wholebody (Jin et al., 2020; Xu et al., 2022a) validation set contains many instances of occluded ears (the ear keypoints are zeroed out in the ground truth annotations), making it challenging to estimate the head keypoint. For the benefit of the Multi-HMR methods, we prune the COCO validation split to only include subjects with visible ear keypoints.

## A.4 Ablation

Our final model requires significant GPU resources to train, so we perform the ablation with earlier checkpoints. In particular, we train the single backbone network's head parameters with the same specifics as described in Section A.2 for 190k iterations and employ a comparable version of our model.

| Models | AGORA | | | | ARCTIC | | | | COCO | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | MPJPE ↓ | PA-MPJPE↓ | @0.05↑ | @0.1↑ | @0.05↑ | @0.1↑ |
| Single Backbone | **141.7** | **60.8** | **0.801** | **0.917** | **101.1** | **46.9** | **0.794** | **0.962** | **0.775** | **0.947** |
| BodhaHMR | 143.3 | 61.0 | 0.797 | 0.916 | 107.6 | 50.8 | 0.775 | 0.947 | 0.744 | 0.923 |

Table 4: **Body results for ablation on hand backbone.**

In Table 4, we report the body results from our ablation study. While the single backbone version consistently outperforms BodhaHMR in this setting, our approach does not lag very far behind. For instance, BodhaHMR is .1% worse on PCK @0.05 for AGORA and 3% worse for COCO. Considering the hand-related improvements we observe in Section 4.4, our method sacrifices minimal body performance to deliver state-of-the-art hand estimations.