MEGA: Masked Generative Autoencoder for Human Mesh Recovery

Guénolé Fiche
CentraleSupélec
IETR UMR CNRS 6164
Cesson-Sévigné, France
guenole.fiche@centralesupelec.fr

Simon Leglaive
CentraleSupélec
IETR UMR CNRS 6164
Cesson-Sévigné, France
simon.leglaive@centralesupelec.fr

Xavier Alameda-Pineda
Inria
UGA, CNRS, LJK
Montbonnot-Saint-Martin, France
xavier.alameda-pineda@inria.fr

Francesc Moreno-Noguer
Institut de Robòtica i Informàtica Industrial
CSIC-UPC
Barcelona, Spain
francesc.moreno.noguer@upc.edu

Abstract

Human Mesh Recovery (HMR) from a single RGB image is a highly ambiguous problem, as similar 2D projections can correspond to multiple 3D interpretations. Nevertheless, most HMR methods overlook this ambiguity and make a single prediction without accounting for the associated uncertainty. A few approaches generate a distribution of human meshes, enabling the sampling of multiple predictions; however, none of them is competitive with the latest single-output model when making a single prediction. This work proposes a new approach based on masked generative modeling. By tokenizing the human pose and shape, we formulate the HMR task as generating a sequence of discrete tokens conditioned on an input image. We introduce MEGA, a MaskEd Generative Autoencoder trained to recover human meshes from images and partial human mesh token sequences. Given an image, our flexible generation scheme allows us to predict a single human mesh in deterministic mode or to generate multiple human meshes in stochastic mode. MEGA enables us to propose multiple outputs and to evaluate the uncertainty of the predictions. Experiments on in-the-wild benchmarks show that MEGA achieves state-of-the-art performance in deterministic and stochastic modes, outperforming single-output and multi-output approaches.

1 Introduction

Perceiving humans from images is a long-standing problem in computer vision, with applications in diverse fields such as sports [20, 78] or e-commerce [51, 85]. Many approaches rely on statistical body models like SMPL [52] for representing humans. Earlier human mesh recovery (HMR) methods recovered the SMPL pose and shape parameters from 2D cues using optimization-based techniques [6, 43]. However, these optimization procedures require good initialization, are time-consuming, and often converge to suboptimal minima. With the advancement of deep learning and the availability of datasets of images with 3D human pose and shape annotations, most approaches have shifted to a regression-based paradigm. Early approaches used architectures based on convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) [34, 46]. Recent works have adopted Transformers [76] for extracting image features or making predictions [11, 24]. Despite achieving unprecedented accuracy, state-of-the-art HMR models still have some weaknesses, such as producing

Preprint. Under review.

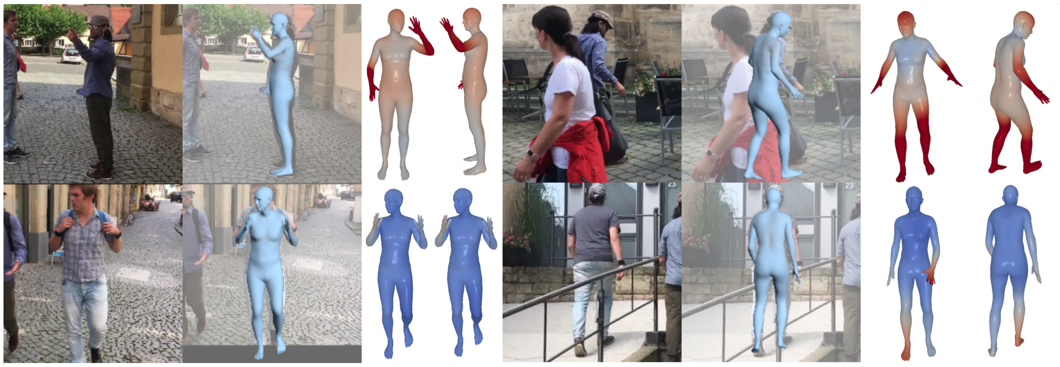

Figure 1: Human mesh recovery from a single image is an ill-posed problem due to depth ambiguity. Probabilistic approaches have aimed to address this by generating multiple predictions, but diversity often sacrifices accuracy. Introducing MEGA, our HMR model based on masked generative modeling achieves state-of-the-art performance on in-the-wild benchmarks while also quantifying uncertainty. Bluish regions in the mesh indicate low uncertainty, while reddish areas signify higher uncertainty.

unrealistic predictions, especially when dealing with occlusions. To address these issues, recent works have proposed tokenizing the human pose using vector quantized variational autoencoders (VQ-VAEs) [19, 21]. This approach uses a discrete representation of the human mesh, learned from large-scale motion capture datasets, to confine predictions to the space of anthropomorphic meshes using a dictionary of valid mesh tokens. This tokenization aligns well with Transformer-based architectures, which were originally designed for processing discrete data in natural language processing. Notably, VQ-HPS [21] reframed HMR as a classification task and achieved state-of-the-art, demonstrating the great potential of human mesh tokenized representations in HMR.

While significant progress has been made in HMR, a major issue remains unaddressed in most prior works: a 2D image, especially with occlusions, cannot provide sufficient information to estimate a 3D human mesh with certainty [57, 71]. This limitation causes single-output models to be biased toward most common poses and body shapes [14]. To mitigate this problem, several works have proposed probabilistic approaches that generate multiple predictions from a single image [41, 68]. These approaches have used various families of generative models, ranging from conditional variational autoencoders (CVAEs) [70] to diffusion models [10]. However, this increase in diversity typically comes at the cost of accuracy [69], and none of these multi-output methods are competitive with the latest single-output HMR models.

In this work, we introduce MEGA, a multi-output HMR approach based on self-supervised learning and masked generative modeling of tokenized human meshes. MEGA relies on the Mesh-VQ-VAE of [21] to encode/decode a 3D human mesh to/from a set of discrete tokens. Our training process unfolds in two steps: (1) Firstly, akin to (vector quantized) masked autoencoders [3, 27, 64, 65], we pre-train MEGA in a self-supervised manner to reconstruct human mesh tokens from partially visible inputs. This leverages large amounts of motion capture data without the need for paired image data. (2) Subsequently, for HMR from RGB images, we train MEGA to predict randomly masked human mesh tokens conditioned on image feature embeddings. During inference, we begin with a fully masked sequence of tokens and generate a human mesh conditioned on an input image. We propose two distinct generation modes: (2.a) In deterministic mode, MEGA predicts all tokens in a single forward pass, ensuring speed and accuracy; (2.b) In stochastic mode, the generation process involves iteratively sampling human mesh tokens, enabling MEGA to produce multiple predictions from a single image for uncertainty quantification.

We evaluate MEGA on in-the-wild HMR benchmarks, comparing it to single-output and probabilistic HMR methods. MEGA achieves state-of-the-art (SOTA) performance when predicting a human mesh in a single forward pass in deterministic mode. In stochastic mode, MEGA outperforms both deterministic and probabilistic methods with a single prediction and significantly enhances its performance with an increased number of samples. This mode generates diverse, realistic human meshes, allowing for uncertainty evaluation and the proposal of multiple plausible outputs given an image. Additionally, we demonstrate that the pre-trained MEGA can generate realistic and diverse random meshes, further showcasing the flexibility of our approach.

In summary, we make the following two key contributions[1]:

- We introduce MEGA, a masked generative autoencoder for human mesh recovery, pre-trained in a self-supervised manner on motion capture data.
- Our flexible inference procedure can operate in deterministic or stochastic modes, with or without image conditioning, achieving state-of-the-art results in all tested scenarios.

2 Related work

2.1 Human mesh recovery

Single output HMR. Since the release of the HMR [34] model, most approaches for recovering human meshes from images have been regression-based, using neural networks to make predictions directly from the image. These regression-based HMR methods can be categorized into parametric and non-parametric approaches. Parametric methods aim to recover the parameters of the SMPL model [19,24,33,36,37,39,45,46,72,83]. They typically produce realistic predictions; however, some works have argued that the SMPL model parameter space is not the most suitable for predicting human meshes [12,14,40], leading to the development of non-parametric approaches. Non-parametric methods predict the coordinates of 3D vertices without relying on a learned parametric model. While earlier approaches used graph convolutional neural network architectures inspired by the mesh topology [40,48], recent non-parametric models predominantly employ Transformers [11, 18, 21, 47, 53]. Although non-parametric methods yield accurate results, they can sometimes produce non-anthropomorphic meshes, particularly when training data is scarce [21].

In this work, we approach HMR differently by formulating it as a classification task. Instead of predicting SMPL model parameters or 3D coordinates, we aim to recover sequences of indices that can be decoded into a human mesh using the Mesh-VQ-VAE from [21].

Multi-output HMR. Estimating a 3D human mesh from a single image is challenging due to the depth ambiguity, especially when the person is partially occluded. Several works have proposed making multiple predictions to address the ill-posed nature of the problem. Earlier works employed compositional models [32] or mixture density networks [44]. More recent approaches rely on sophisticated probabilistic distributions [67,68] and generative models, such as CVAE [70], normalizing flows [4, 41, 69], and diffusion models [10]. While these methods can predict diverse plausible solutions, they often face a trade-off between accuracy and diversity [69] and struggle to achieve SOTA results even when making multiple predictions.

MEGA is based on masked generative modeling to produce multiple predictions. Our experiments demonstrate that, while generating diverse samples, MEGA outperforms SOTA approaches even with a single prediction in stochastic mode and significantly improves as the number of samples increases.

2.2 Self-supervised learning for HMR

Self-supervised learning (SSL) approaches can be categorized into two families: discriminative and generative [50, 59, 82]. Many prior works used discriminative SSL approaches to train 3D human pose estimation models. Most of these methods exploit multi-view consistency constraints for supervision [9,38,63,77], while others use temporal consistency in videos [42,66,74] or images with different resolutions [81]. [13] explored the use of discriminative SSL for pre-training human mesh estimator backbones, demonstrating that 2D annotation-based pre-training leads to faster convergence and improved results. However, [2] surpassed traditional feature extractors by employing generative SSL, using cross-view and cross-pose completion to train a Vision Transformer (ViT) [17].

While prior works have demonstrated the importance of pre-training for the backbone of HMR models [5, 60], we propose pre-training the generative model on human meshes to leverage extensive motion capture data. This pre-training not only provides us with an unconditional human mesh generative model but also proves beneficial for training a robust HMR model. In deterministic mode,

[1] Code and trained models will be released upon acceptance.

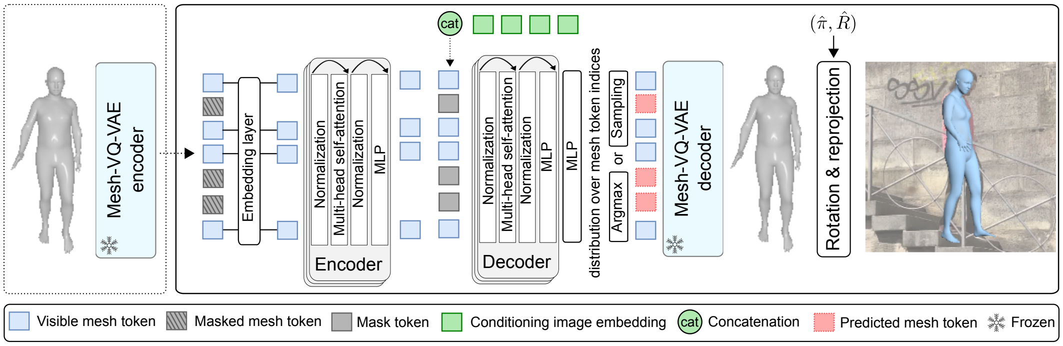

Figure 2: MEGA is a masked generative model based on an encoder-decoder Transformer architecture. During the self-supervised pretraining stage, MEGA is trained to predict human mesh tokens from partially visible inputs using motion capture data without paired image data. During the supervised training stage for HMR, the model is trained to predict randomly masked human mesh tokens conditioned on image embeddings. For both training stages, only the cross-entropy loss is used on the predicted mesh tokens. At test time, in stochastic inference mode, we start from a fully masked sequence of tokens and iteratively sample human mesh tokens conditioned on input image embeddings. In deterministic inference mode, we predict all tokens in a single forward pass.

the encoder of our model can be discarded, as the decoder alone is sufficient for predicting sequences in a single forward pass (see Sec. 3.4). To our knowledge, this is the first work that discards the encoder of an MAE after pre-training; previous works typically discard the decoder and use the encoder to extract features.

2.3 Masked generative modeling

Masked modeling was introduced in BERT [16] for language modeling and extended to images with the MAE [27]. This technique trains a model to predict randomly masked tokens in a sequence based on visible tokens. Masked generative modeling builds on this by training a model to generate new samples, starting from a fully masked sequence and iteratively predicting a fixed number of tokens at each step [7, 8].

In this work, we develop a masked generative modeling approach for HMR. Using a tokenized representation of the human mesh is particularly well-suited for this task, as it allows for straightforward masking and replacement of mesh parts with mask tokens.

3 MEGA

3.1 Human mesh tokenization

MEGA relies on a tokenized representation of the human mesh. Specifically, we use the Mesh-VQ-VAE introduced in [21], which is a VQ-VAE [75] with a fully convolutional mesh autoencoder architecture [86]. The Mesh-VQ-VAE tokenizes canonical human meshes following the SMPL topology, with zero translation and facing the camera. The input canonical mesh with vertices $V_c \in \mathbb{R}^{6890 \times 3}$ is encoded into a sequence of $N = 54$ latent vectors, each of dimension $L = 9$. As the architecture is fully convolutional, each latent vector encodes a specific part of the human body. Through vector quantization, each latent vector is replaced by a human mesh token index corresponding to an embedding vector of dimension L in a codebook of size $S = 512$. Thus, a human mesh is represented by a sequence of N token indices in $\{1, ..., S\}$, which can be decoded with the Mesh-VQ-VAE decoder to reconstruct the vertices $\hat{V}_c$. The Mesh-VQ-VAE is pre-trained on motion capture data and remains frozen during MEGA's training. In this paper, we formulate HMR as the task of generating human mesh tokens conditioned on an input image.

3.2 Model

Overall pipeline. MEGA is a masked generative model based on an encoder-decoder Transformer architecture, illustrated in Fig. 2. During training, the Mesh-VQ-VAE encoder converts a 3D human mesh into a sequence of N tokens. These human mesh tokens are then randomly masked, leaving only $M < N$ visible tokens. An embedding layer converts the visible token indices into learned token embeddings, which are subsequently passed to an encoder with multi-head self-attention. The sequence of encoded token embeddings is completed with mask tokens and, during supervised training and inference stages only, it is also concatenated with a sequence of image embeddings (see next paragraph). The complete sequence of tokens is then processed by a decoder with multi-head self-attention, followed by an MLP that outputs a distribution over the N human mesh token indices (in practice, it outputs the logits, which can be normalized using a softmax function). During training and in deterministic inference mode, the $N - M$ human mesh tokens that were originally masked are predicted by taking the argmax, and the Mesh-VQ-VAE decoder is used to reconstruct the canonical mesh $\hat{V}_c$. A simple cross-entropy loss is computed from the predicted mesh tokens for both the self-supervised and supervised training stages. In stochastic inference mode, the human mesh tokens are predicted iteratively by sampling the predicted distribution over the mesh token indices and reintroducing a proportion of unmasked tokens at the encoder's input. Finally, an MLP (not represented in Fig. 2) is trained to predict a global 6D rotation $\hat{R} \in \mathbb{R}^6$ and the perspective camera parameters $\hat{\pi} \in \mathbb{R}^3$ from image features. These can be used to orient the predicted canonical mesh and to reproject it on the input image.

Image embeddings extractor. An image features extractor computes image features $X \in \mathbb{R}^{WH \times C}$. For fair comparisons with SOTA methods, we rely on an HRNet [79] backbone ($W = H = 7$ and $C = 720$) pre-trained on a 2D pose estimation task [49]. Additionally, we provide results using a more powerful ViT [17] backbone ($W = 12$, $H = 16$, $C = 1280$). For the multi-output HMR experiment, we adhere to the common practice in the literature and use a ResNet-50 [28] backbone ($W = H = 7$, $C = 2048$). Prior to being fed to the MEGA decoder, the image features are linearly projected to the Transformer dimension, resulting in WH image embeddings of dimension $D = 1024$.

Encoder. The encoder of MEGA consists of $B_e = 12$ blocks, akin to the Vision Transformer [17]. Each block contains a multi-head self-attention and an MLP module, with layer normalization preceding and residual connections following every module (see Fig. 2). To preserve positional information at the input of the encoder, learned position embeddings [22] are added to the human mesh token embeddings, and a *cls* token [16] is concatenated with the input sequence.

Decoder. The decoder of MEGA consists of $B_d = 4$ blocks identical to the encoder blocks. During the self-supervised pre-training on human meshes (see Sec. 3.3), the decoder receives only the sequence of encoded human mesh token embeddings, completed with mask tokens (a single trainable embedding vector repeated at each masked token position), in line with the masked autoencoder strategy [27]. During the supervised training stage and at inference, this input sequence is concatenated with the sequence of image embeddings. Position embeddings [22, 76] are added at the input of the decoder.

Rotation and camera prediction. We use the previously mentioned image features $X \in \mathbb{R}^{WH \times C}$ to predict the global 6D rotation $\hat{R} \in \mathbb{R}^6$ and the perspective camera parameters $\hat{\pi} \in \mathbb{R}^3$. The WH image features are averaged to yield a single vector of dimension C, which is subsequently passed through an MLP with 2 hidden layers. The output of this MLP is then linearly mapped to the rotation and camera parameters.

3.3 Training strategy

Self-supervised pre-training. MEGA is pre-trained in a self-supervised manner on tokenized human meshes using a strategy similar to vector quantized masked autoencoders [3, 27, 64, 65]. The pre-training task involves reconstructing randomly masked human mesh tokens from a set of visible tokens. A variable masking rate is used such that $M = \lfloor N \cos(\frac{\pi\tau}{2}) \rfloor$ with τ uniformly sampled from $[0, 1[$. The variable masking rate is critical for allowing MEGA to generate meshes iteratively in stochastic mode, as each step of the generation process involves predicting all tokens given a variable number of visible tokens (see Sec. 3.4). The sole loss used for pre-training is a cross-entropy loss computed from the reconstruction of the masked tokens.

Masked generative modeling for HMR. To train MEGA to predict tokenized canonical meshes from images, we extend the pre-training strategy by conditioning the decoder with the image embeddings (see Sec. 3.2 and Fig. 2). The human mesh tokens follow the same masking rate schedule, while the image embeddings remain fully visible. The only supervision for predicting canonical meshes is a cross-entropy loss, as in the pre-training stage. For predicting the rotation and camera parameters, we use the Euclidean norm on the rotation matrix corresponding to the predicted 6D representation and an $L1$ loss on the reprojection of 2D joints extracted from the predicted oriented mesh, using the predicted perspective camera parameters.

3.4 Generation strategy

We propose two generation modes for inference: deterministic and stochastic. Both modes start from a fully masked sequence of human mesh tokens, aiming to generate a complete sequence that can be decoded into a canonical mesh. Regardless of the generation mode, the camera and rotation are predicted deterministically from the image.

Deterministic mode. In deterministic inference mode, we predict all tokens in a single forward pass by taking the argmax of the predicted distribution over the human mesh token indices. In this mode, MEGA's encoder is not used; instead, the decoder is fed with a sequence of N mask tokens, relying entirely on the image representation information. Consequently, the encoder can be discarded, significantly reducing the model size as $B_e > B_d$. To our knowledge, this is the first work to discard the encoder of an MAE, whereas previous works typically discard the decoder and use the encoder to obtain representations for downstream tasks.

Stochastic mode. MEGA addresses the ambiguity of the HMR from a single image by operating in a stochastic inference mode and generating diverse plausible human meshes. We follow a strategy similar to [7, 8, 25], employing an iterative generation process in T steps. At each step $t \in \{1, ..., T\}$, we predict $n_t - n_{t-1}$ tokens, where $n_t = \lfloor N \times (1 - \cos(\frac{\pi t}{2T}))\rfloor$ denotes the number of predicted tokens up to step t. The tokens predicted at a given step remain visible for the subsequent steps. By the end of this iterative process, we generate $n_T = N$ tokens that represent a complete human mesh. The prediction at step $t \in \{1, ..., T\}$ proceeds as follows:

- The currently visible n_{t-1} tokens are fed to the model.
- For each of the $P_t = N - n_{t-1}$ tokens that are still masked, the model outputs unnormalized probabilities over the indices of the Mesh-VQ-VAE codebook. These probabilities are sampled using the Gumble-max trick [30] to obtain P_t candidate tokens to be set visible in the next step. Each candidate token is identified by an index between 1 and S (the codebook size) along with its unnormalized probability.
- Using the Gumble-max trick again, we finally sample $n_t - n_{t-1}$ tokens among the set of P_t candidate tokens, which will be visible in the next step.

Given the stochastic nature of this generation mode, we can obtain Q different human mesh predictions for a single image by repeating the above generation process several times. The uncertainty of each vertex can then be computed as its standard deviation over the Q predictions.

4 Experiments

4.1 Experimental setup

Datasets. MEGA is initially pre-trained in a self-supervised manner on a diverse subset of the AMASS dataset [54], focusing on samples with high pose and body shape variety, as detailed in [5]. MEGA is then trained for HMR using a mix of standard image datasets labeled with human meshes, including MSCOCO [49], Human3.6M [31], MPI-INF-3DHP [56] and MPII [1]. We evaluate MEGA and compare it to the SOTA methods on the in-the-wild 3DPW [55] and EMDB [35] datasets. Following recent works [19, 24], we do not finetune MEGA on the 3DPW training set before evaluation. This approach better validates the model's generalization capacity and allows us to use the same model for all experiments. Unless specified otherwise, other models in the comparison tables use the same training datasets and the same image feature extractor as ours for fair comparisons.

Table 1: **Evaluation in deterministic mode.** We evaluate MEGA on the 3DPW and EMDB datasets and compare it to the SOTA methods using metrics defined in Sec. 4.1 given in mm. † stands for additionally using 2D training data, and ∓ for additionally using 2D data and Bedlam [5].

		3DPW			EMDB		
Method	Backbone	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓
FastMETRO [11]	HRNet-w64	121.6	109.0	65.7	119.2	108.1	72.7
PARE [37]	HRNet-w32	97.9	82.0	50.9	133.2	113.9	72.2
Virtual Marker [53]	HRNet-w48	93.8	80.5	48.9	-	-	-
CLIFF [46]	HRNet-w48	87.6	73.9	46.4	122.9	103.1	68.8
VQ-HPS [21]	HRNet-w48	84.8	71.1	45.2	112.9	99.9	65.2
HMR2.0 [24]	ViT-H	84.1	70.0	44.5	120.1	97.8	61.5
TokenHMR† [19]	ViT-H	88.1	76.2	49.3	124.4	102.4	67.5
TokenHMR$^\mp$ [19]	ViT-H	84.6	71.0	44.3	109.4	91.7	55.6
MEGA w/o pre-training (ours)	HRNet-w48	84.1	70.5	46.2	113.9	95.9	62.0
MEGA (ours)	HRNet-w48	81.6	68.5	44.1	**107.9**	**90.5**	58.7
MEGA (ours)	ViT-H	**80.0**	**67.5**	**41.0**	108.6	92.4	**52.5**

Table 2: **Evaluation in stochastic mode.** We compare MEGA to the SOTA probabilistic methods on the multi-output HMR task using standard metrics (see Sec. 4.1) given in mm and the relative improvement (Imp) in %. ‡ uses an HRNet backbone; all other methods use a ResNet-50 backbone.

Method	PVE ↓				Imp ↑	MPJPE ↓				Imp ↑	PA-MPJPE ↓				Imp ↑
Q	1	5	10	25		1	5	10	25		1	5	10	25	
Diff-HMR [10]	114.6	111.8	110.9	109.8	4.2	98.9	96.3	95.5	94.5	4.5	58.5	57.0	56.5	55.9	4.4
3D Multibodies [4]	-	-	-	-	-	93.8	82.2	79.4	75.8	**19.2**	59.9	57.1	56.6	55.6	7.2
ProHMR [41]	-	-	-	-	-	97.0	93.1	89.8	84.0	13.4	59.8	56.5	54.6	52.4	12.4
MEGA (ours)	101.6	92.8	90.4	87.5	**13.9**	86.2	78.0	76.4	73.9	14.3	58.6	51.6	49.7	47.6	**18.7**
MEGA det (ours)	90.6					76.3					48.3				
MEGA‡ (ours)	**83.4**	**78.4**	**76.9**	**75.1**	10.0	**69.9**	**66.1**	**64.9**	**63.6**	9.0	**45.5**	**42.6**	**41.7**	**40.4**	11.2

Implementation details. All experiments are conducted using PyTorch [61]. MEGA is pre-trained on human meshes for 500 epochs. For the HMR task, MEGA is first trained on MSCOCO [49] for 100 epochs, followed by training on the mix of datasets described above for 10 epochs. Using 4 NVIDIA A100 GPUs, the entire pre-training and training process takes approximately 3 days.

Metrics. To evaluate HMR methods, we use the widely adopted metrics: per-vertex-error (PVE), mean-per-vertex-error (MPJPE), and the Procrustes-aligned MPJPE (PA-MPJPE). PVE measures the Euclidean distance between the vertices of the predicted mesh and the ground truth. MPJPE assesses the accuracy of body joints extracted from the mesh. PA-MPJPE is similar to MPJPE but includes a Procrustes-alignment, a rigid transformation that minimizes the distance between the predicted and ground truth joints. All three metrics are reported in mm.

4.2 Deterministic inference mode

We evaluate MEGA on the HMR task using the deterministic inference mode defined in Sec. 3.4, and compare its performance to SOTA methods on 3DPW [55] and EMDB [35]. Results for other methods are taken from the corresponding papers when available or computed using official implementations and model weights. To ensure fairness, we only compare methods trained on standard datasets (see Sec. 4.1), with the exception of TokenHMR [19], which uses additional data but is included in Tab. 1 because it also uses mesh tokenization. Note that the results for FastMETRO [11] and Virtual Marker [53] differ from those in their original papers because we present results without finetuning on the 3DPW dataset. Non-parametric methods predicting 3D coordinates often struggle to generalize to unseen datasets, and skipping finetuning on the 3DPW training set significantly degrades their performance.

MEGA outperforms all other methods on in-the-wild datasets. Using an HRNet [79] backbone, MEGA significantly surpasses both parametric [37, 46] and non-parametric [11, 21, 53] methods, especially on the EMDB [35] dataset. With a ViT [17] backbone, MEGA also achieves SOTA performance. However, comparisons with TokenHMR [19] are not completely fair, as this method is trained with additional 2D data and Bedlam [5]. We also provide results of MEGA without pre-training on motion capture data, demonstrating the interest of the self-supervised pre-training step.

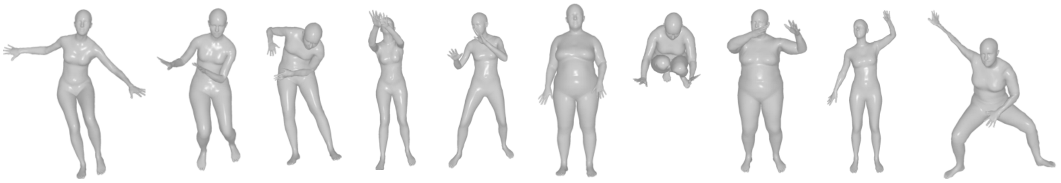

Figure 3: **Random mesh generations.** We use MEGA pre-trained in a self-supervised fashion to generate random human meshes.

4.3 Stochastic inference mode

We evaluate MEGA in stochastic inference mode (see Sec. 3.4) for the HMR task on the 3DPW [55] dataset, comparing our performance with SOTA multi-output approaches. Results from other methods are obtained from their respective papers. We exclusively compare methods trained on standard datasets (see Sec. 4.1), providing a version of MEGA using a ResNet-50 [28] backbone for fair comparisons. Following prior works [4, 10, 41], we assess accuracy using standard metrics computed with the minimum error sample out of Q predictions, with Q ranging from 1 to 25. We also calculate the relative improvement between 1 and 25 samples. Results are presented in Tab. 2.

MEGA achieves SOTA performance across all metrics and sample sizes. Notably, MEGA exhibits significantly higher accuracy than other methods when generating a single sample, and consistently demonstrates the best or second-best relative improvement among methods using a ResNet-50 backbone. With an HRNet [79] backbone, MEGA outperforms all probabilistic methods and surpasses SOTA single-output methods, even with a single stochastic generation. It is worth noting that it takes between 10 and 25 stochastic samples to outperform the deterministic generation (MEGA det), highlighting the utility of deterministic mode for quick and accurate predictions and the advantage of multiple predictions for enhanced accuracy. For a detailed comparison between deterministic and stochastic modes, please refer to Appendix B.

4.4 Random meshes generation

We propose to use MEGA pre-trained in a self-supervised manner (see Sec. 3.3) for generating random human meshes. For comparison, we assess its capabilities against VPoser [62] and NRDF [29]. VPoser is a conventional pose prior in VAE form, while NRDF is a SOTA pose prior based on neural fields [80]. Although these models do not explicitly model body shapes, making them not directly comparable to MEGA, which directly generates meshes with diverse poses and shapes, they are the most suitable for comparison purposes. As far as we know, MEGA is the first model generating unconditioned random human body meshes with pose and shape diversity. We assess all 3 models in terms of diversity using the average pairwise distance (APD) in cm, representing the average distance between the joints of all pairs of samples. For plausibility evaluation, we compute the Fréchet inception distance (FID) with the fully convolutional mesh autoencoder introduced in [86] trained on AMASS [54], with a latent space dimension of 7×9. The FID compares the latent representation of generated meshes with that of a representative subset of AMASS introduced in [5].

We randomly sample 500 meshes with each method. Regarding plausibility, MEGA outperforms other methods, achieving an FID of 0.001 compared to 0.007 for VPoser and 0.033 for NRDF. This result is not surprising, as other methods use the average shape for all meshes, whereas MEGA produces diverse results in poses and shapes. NRDF generates more diverse meshes, with an APD of 28.61 cm, while VPoser and MEGA achieve APDs of 18.32 and 20.77 cm, respectively. In summary, MEGA clearly outperforms VPoser, as our generated samples are more diverse and plausible. NRDF produces more diverse poses, but the distribution of the generation samples of MEGA is more representative of the AMASS dataset. In Fig. 3, we present some qualitative samples of MEGA's generation, which exhibit diverse and realistic poses and shapes.

4.5 Stochastic inference mode analysis

We evaluate the correlation between the PVE in deterministic mode and the uncertainty of the predictions in stochastic mode, quantified by the standard deviation of the predicted vertices. We conduct this experiment on 3DPW-OCC [84], which features numerous occlusions, thereby increasing prediction uncertainty. Fig. 5 features multiple samples for images with occlusions, accompanied

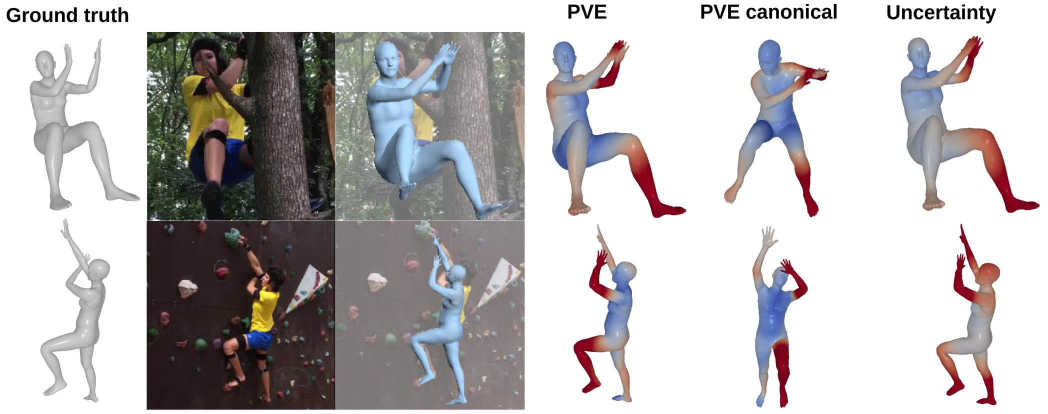

Figure 4: **PVE and uncertainty.** We visualize the PVE and the standard deviation in colormaps on meshes.

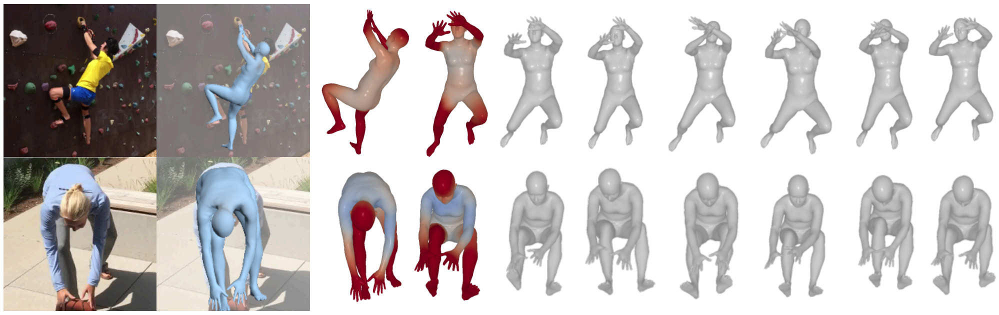

Figure 5: **Multiple predictions given an image.** We generate multiple human meshes given images with occlusions [84] using MEGA in stochastic mode. Note that, while not explicitly shown in the figure, all of these meshes practically have identical image projections.

by a quantitative evaluation in deterministic mode on 3DPW-OCC in Appendix A. In Fig. 4, we visually represent the PVE and the uncertainty as colormaps on meshes. Interestingly, we observe a striking similarity between them, suggesting that the estimated uncertainty could be used to evaluate the model's per-vertex performance.

5 Conclusion

In this work, we explored self-supervised learning and masked generative modeling on human meshes for HMR. We introduced MEGA, a masked generative autoencoder designed to generate human meshes as discrete token sequences. MEGA's flexible architecture and generation scheme enable the generation of diverse and realistic meshes and support both single and multi-output HMR. A thorough evaluation demonstrates significant improvement over SOTA in both these domains.

Limitations. While MEGA generally produces accurate predictions, it struggles with extreme poses significantly divergent from the training data. Although uncertainty estimation shows promise, further analysis is required before integrating them into model training supervision. Fig. 8 in the Appendix provides visualizations of these failure cases. Additionally, the experiments were conducted only once, without testing different seeds to confirm the results' significance. However, it allowed us to significantly reduce our research's environmental impact.

Future work. MEGA's adaptable framework suggests potential applications beyond its current scope. Future research could explore generating human meshes conditioned on text inputs [15]. We could also complement image embedding with more observations, such as 2D pose tokens [23] or tokenized meshes of other individuals to model social interactions [26, 58]. Extending this work to videos by incorporating temporal masking during training [64, 65] or including more extreme poses in training data [73] may further improve performance. Additionally, optimizing MEGA's computational efficiency is a promising avenue for future research [18].

Broader impact. MEGA contributes to the understanding of human perception from images. While there's concern about potential misuse for intrusive surveillance, MEGA doesn't reconstruct facial features, preserving anonymity. MEGA could have applications in healthcare, such as motor assessment of patients. This application would be positive, but potential prediction errors could negatively affect the care pathway. As with all deep learning models, training MEGA demands significant computing power, raising environmental concerns. Hyperparameter optimization wasn't prioritized to minimize this cost, with model training taking about 9 days in total on 4 A100 GPUs for reported experiments.

Acknowledgments

This study is part of the EUR DIGISPORT project supported by the ANR within the framework of the PIA France 2030 (ANR-18-EURE-0022). This work was performed using HPC resources from the "Mésocentre" computing center of CentraleSupélec, École Normale Supérieure Paris-Saclay, and Université Paris-Saclay supported by CNRS and Région Île-de-France. We also thank Samir Sadok for his valuable help with the implementation of the masked autoencoder.

References

[1] Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3686–3693 (2014)

[2] Armando, M., Galaaoui, S., Baradel, F., Lucas, T., Leroy, V., Brégier, R., Weinzaepfel, P., Rogez, G.: Cross-view and cross-pose completion for 3D human understanding. arXiv preprint arXiv:2311.09104 (2023)

[3] Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (ICLR) (2022)

[4] Biggs, B., Novotny, D., Ehrhardt, S., Joo, H., Graham, B., Vedaldi, A.: 3D Multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 20496–20507 (2020)

[5] Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737 (2023)

[6] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: European Conference on Computer Vision (ECCV). pp. 561–578. Springer (2016)

[7] Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)

[8] Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: MaskGIT: Masked generative image transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11315–11325 (2022)

[9] Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Mv, R., Stojanov, S., Rehg, J.M.: Unsupervised 3D pose estimation with geometric self-supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5714–5724 (2019)

[10] Cho, H., Kim, J.: Generative approach for probabilistic human mesh recovery using diffusion models. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4183–4188 (2023)

[11] Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In: European Conference on Computer Vision (ECCV). pp. 342–359. Springer (2022)

[12] Choi, H., Moon, G., Lee, K.M.: Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In: European Conference on Computer Vision (ECCV). pp. 769–787. Springer (2020)

[13] Choi, H., Nam, H., Lee, T., Moon, G., Lee, K.M.: Rethinking self-supervised visual representation learning in pre-training for 3D human pose and shape estimation. In: International Conference on Learning Representations (ICLR) (2022)

[14] Corona, E., Pons-Moll, G., Alenyà, G., Moreno-Noguer, F.: Learned vertex descent: a new direction for 3D human model fitting. In: European Conference on Computer Vision (ECCV). pp. 146–165. Springer (2022)

[15] Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: PoseScript: 3D human poses from natural language. In: European Conference on Computer Vision (ECCV). pp. 346–362. Springer (2022)

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[17] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

[18] Dou, Z., Wu, Q., Lin, C., Cao, Z., Wu, Q., Wan, W., Komura, T., Wang, W.: TORE: Token reduction for efficient human mesh recovery with transformer. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15143–15155 (2023)

[19] Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: TokenHMR: Advancing human mesh recovery with a tokenized pose representation. arXiv preprint arXiv:2404.16752 (2024)

[20] Einfalt, M., Zecha, D., Lienhart, R.: Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming. In: IEEE/CVF Winter conference on Applications of Computer Vision (WACV). pp. 446–455. IEEE (2018)

[21] Fiche, G., Leglaive, S., Alameda-Pineda, X., Agudo, A., Moreno-Noguer, F.: VQ-HPS: Human pose and shape estimation in a vector-quantized latent space. arXiv preprint arXiv:2312.08291 (2023)

[22] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International conference on machine learning (ICML). pp. 1243–1252. PMLR (2017)

[23] Geng, Z., Wang, C., Wei, Y., Liu, Z., Li, H., Hu, H.: Human pose as compositional tokens. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 660–671 (2023)

[24] Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14783–14794 (2023)

[25] Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: MoMask: Generative masked modeling of 3D human motions. arXiv preprint arXiv:2312.00063 (2023)

[26] Guo, W., Bie, X., Alameda-Pineda, X., Moreno-Noguer, F.: Multi-person extreme motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13053–13064 (2022)

[27] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)

[28] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

[29] He, Y., Tiwari, G., Birdal, T., Lenssen, J.E., Pons-Moll, G.: NRDF: Neural riemannian distance fields for learning articulated pose priors. arXiv preprint arXiv:2403.03122 (2024)

[30] Huijben, I.A., Kool, W., Paulus, M.B., Van Sloun, R.J.: A review of the Gumbel-max trick and its extensions for discrete stochasticity in machine learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **45**(2), 1353–1371 (2022)

[31] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36**(7), 1325–1339 (2013)

[32] Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 805–814 (2017)

[33] Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3d human pose estimation. In: International Conference on 3D Vision (3DV). pp. 42–52. IEEE (2021)

[34] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)

[35] Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14632–14643 (2023)

[36] Khirodkar, R., Tripathi, S., Kitani, K.: Occluded human mesh recovery. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1715–1725 (2022)

[37] Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11127–11137 (2021)

[38] Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3D human pose using multi-view geometry. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1077–1086 (2019)

[39] Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019)

[40] Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4501–4510 (2019)

[41] Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11605–11614 (2021)

[42] Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3D human pose estimation via part guided novel image synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6152–6162 (2020)

[43] Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3D and 2D human representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6050–6059 (2017)

[44] Li, C., Lee, G.H.: Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9887–9895 (2019)

[45] Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3383–3393 (2021)

[46] Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: CLIFF: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision (ECCV). pp. 590–606. Springer (2022)

[47] Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1954–1963 (2021)

[48] Lin, K., Wang, L., Liu, Z.: Mesh Graphormer. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12939–12948 (2021)

[49] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)

[50] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering **35**(1), 857–876 (2021)

[51] Liu, Y., Liu, Y., Xu, S., Cheng, K., Masuko, S., Tanaka, J.: Comparing vr-and ar-based try-on systems using personalized avatars. Electronics **9**(11), 1814 (2020)

[52] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015)

[53] Ma, X., Su, J., Wang, C., Zhu, W., Wang, Y.: 3D human mesh estimation from virtual markers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 534–543 (2023)

[54] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5442–5451 (2019)

[55] von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: European Conference on Computer Vision (ECCV). pp. 601–617. Springer (2018)

[56] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: International Conference on 3D Vision (3DV). pp. 506–516. IEEE (2017)

[57] Moreno-Noguer, F., Fua, P.: Stochastic exploration of ambiguities for nonrigid shape recovery. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **35**(2), 463–475 (2012)

[58] Müller, L., Ye, V., Pavlakos, G., Black, M., Kanazawa, A.: Generative proxemics: A prior for 3D social interaction from images. arXiv preprint arXiv:2306.09337 (2023)

[59] Ozbulak, U., Lee, H.J., Boga, B., Anzaku, E.T., Park, H., Van Messem, A., De Neve, W., Vankerschaver, J.: Know your self-supervised learning: A survey on image-based generative and discriminative training. arXiv preprint arXiv:2305.13689 (2023)

[60] Pang, H.E., Cai, Z., Yang, L., Zhang, T., Liu, Z.: Benchmarking and analyzing 3D human pose and shape estimation beyond algorithms. Advances in Neural Information Processing Systems (NeurIPS) **35**, 26034–26051 (2022)

[61] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)

[62] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)

[63] Roy, S.K., Citraro, L., Honari, S., Fua, P.: On triangulation as a form of self-supervision for 3D human pose estimation. In: International Conference on 3D Vision (3DV). pp. 1–10. IEEE (2022)

[64] Sadok, S., Leglaive, S., Séguier, R.: A vector quantized masked autoencoder for audiovisual speech emotion recognition. arXiv preprint arXiv:2305.03568 (2023)

[65] Sadok, S., Leglaive, S., Séguier, R.: A vector quantized masked autoencoder for speech emotion recognition. In: 2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW). pp. 1–5. IEEE (2023)

[66] Schmidtke, L., Hou, B., Vlontzos, A., Kainz, B.: Self-supervised 3D human pose estimation in static video via neural rendering. In: European Conference on Computer Vision (ECCV). pp. 704–713. Springer (2022)

[67] Sengupta, A., Budvytis, I., Cipolla, R.: Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11219–11229 (2021)

[68] Sengupta, A., Budvytis, I., Cipolla, R.: Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16094–16104 (2021)

[69] Sengupta, A., Budvytis, I., Cipolla, R.: HuManiFlow: Ancestor-conditioned normalising flows on SO(3) manifolds for human pose and shape distribution estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4779–4789 (2023)

[70] Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3D human pose estimation by generation and ordinal ranking. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2325–2334 (2019)

[71] Simo-Serra, E., Ramisa, A., Alenya, G., Torras, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2673–2680 (2012)

[72] Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3D people. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11179–11188 (2021)

[73] Tripathi, S., Müller, L., Huang, C.H.P., Taheri, O., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4713–4725 (2023)

[74] Tung, H.Y., Tung, H.W., Yumer, E., Fragkiadaki, K.: Self-supervised learning of motion capture. Advances in Neural Information Processing Systems (NeurIPS) **30**, 5236–5246 (2017)

[75] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in Neural Information Processing Systems (NeurIPS) **30**, 6306–6315 (2017)

[76] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) **30**, 5998–6008 (2017)

[77] Wandt, B., Rudolph, M., Zell, P., Rhodin, H., Rosenhahn, B.: CanonPose: Self-supervised monocular 3D human pose estimation in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13294–13304 (2021)

[78] Wang, J., Qiu, K., Peng, H., Fu, J., Zhu, J.: AI Coach: Deep human pose estimation and analysis for personalized athletic training assistance. In: ACM International Conference on Multimedia (ACM MM). pp. 374–382 (2019)

[79] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI) **43**(10), 3349–3364 (2020)

[80] Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. In: Computer Graphics Forum. vol. 41, pp. 641–676. Wiley Online Library (2022)

[81] Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., De la Torre, F.: 3D human shape and pose from a single low-resolution image with self-supervised learning. In: European Conference on Computer Vision (ECCV). pp. 284–300. Springer (2020)

[82] Zhang, C., Zhang, C., Song, J., Yi, J.S.K., Zhang, K., Kweon, I.S.: A survey on masked autoencoder for self-supervised learning in vision and beyond. arXiv preprint arXiv:2208.00173 (2022)

[83] Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11446–11456 (2021)

[84] Zhang, T., Huang, B., Wang, Y.: Object-occluded human shape and pose estimation from a single color image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7376–7385 (2020)

[85] Zheng, N., Song, X., Chen, Z., Hu, L., Cao, D., Nie, L.: Virtually trying on new clothing with arbitrary poses. In: ACM International Conference on Multimedia (ACM MM). pp. 266–274 (2019)

[86] Zhou, Y., Wu, C., Li, Z., Cao, C., Ye, Y., Saragih, J., Li, H., Sheikh, Y.: Fully convolutional mesh autoencoder using efficient spatially varying kernels. Advances in Neural Information Processing Systems (NeurIPS) **33**, 9251–9262 (2020)

Table 3: **Evaluation on 3DPW-OCC.** We evaluate MEGA on an occlusion dataset and compare it to SOTA occluded HMR methods using standard metrics (see Sec. 4.1) in mm.

		3DPW-OCC		
Method	Backbone	PVE ↓	MPJPE ↓	PA-MPJPE ↓
ROMP [Sun et al., 2021]	ResNet-50	-	-	65.9
SPIN [Kolotouros et al., 2019]	Resnet-50	121.6	95.6	60.8
VisDB [Yao et al., 2022]	ResNet-50	110.5	87.3	56.0
PARE [Kocabas et al., 2021]	HRNet-w32	107.9	90.5	57.1
3DCrowdNet [Choi et al., 2022]	ResNet-50	103.2	88.6	56.8
SEFD [Yang et al., 2023]	ResNet-50	97.1	83.5	55.0
MEGA (ours)	ResNet-50	93.8	79.8	51.5
MEGA (ours)	HRNet-w48	78.9	66.3	43.7

Table 4: **Comparison between deterministic and stochastic generation modes.** In stochastic mode, we evaluate the mean mesh obtained with different sample sizes on 10% of the 3DPW [55] dataset, and we provide its distance to the deterministic prediction (Dist. to det.). We also report the standard deviation of the predictions. All metrics are in mm.

	Stochastic						Deterministic
Q	1	5	10	25	50	100	
PVE ↓	84.08	83.27	83.13	83.09	83.02	83.05	83.10
MPJPE ↓	70.88	70.29	70.15	70.09	70.04	70.05	69.95
PA-MPJPE ↓	44.42	43.77	43.66	43.65	43.62	43.61	43.53
Dist. to det.	14.78	9.53	8.56	7.95	7.71	7.61	0.0
SD	N/A	11.61	12.30	12.65	12.79	12.88	0.0

A Performance on an occlusion dataset

We quantitatively evaluate MEGA on the occlusion dataset 3DPW-OCC [Zhang et al., 2020] in Tab. 3. Despite not being tailored for occluded HMR, MEGA surpasses all other methods trained on the same data (see Sec. 4.1). This performance could stem from MEGA's self-attention mechanism among mesh tokens. While visible parts can be predicted by leveraging image embeddings, occluded parts heavily rely on visible body parts for accurate inference.

B Link between the deterministic and stochastic modes

To gain deeper insights into the stochastic generation mode, we propose not only evaluating the best sample among the Q generations (the common practice in the literature), but also assessing the mean of the generated meshes. In Tab. 4, we compare the performance of the average prediction for different Q using the standard metrics (refer to Sec. 4.1). Additionally, we compute the Euclidean distance between the mean mesh and the one obtained in deterministic mode (Dist. to det., in mm) and the standard deviation of the predictions averaged over all the vertices (SD, in mm). To reduce the computational costs, this study is conducted on a randomly selected 10% subset of images from the 3DPW dataset. As Q increases, the average prediction in stochastic mode approaches the deterministic prediction, with a distance around 7 mm for $Q = 100$. The average prediction appears to converge toward a favorable solution, slightly outperforming the deterministic prediction in terms of PVE.

We also examine the distributions of PVE, MPJPE, and PA-MPJPE on the 3DPW [von Marcard et al., 2018] dataset for MEGA in both deterministic and stochastic modes across various sample sizes Q. In stochastic mode, we analyze the average and best predictions. The results are reported in Fig. 6. Notably, the average prediction of the stochastic mode appears to converge toward the deterministic prediction, particularly as Q increases; the distributions are very similar, with overlapping 95% confidence intervals. This corroborates the findings presented in Tab. 4. When Q equals 1, the mean performance is comparatively lower, resulting in higher error values. These observations underscore

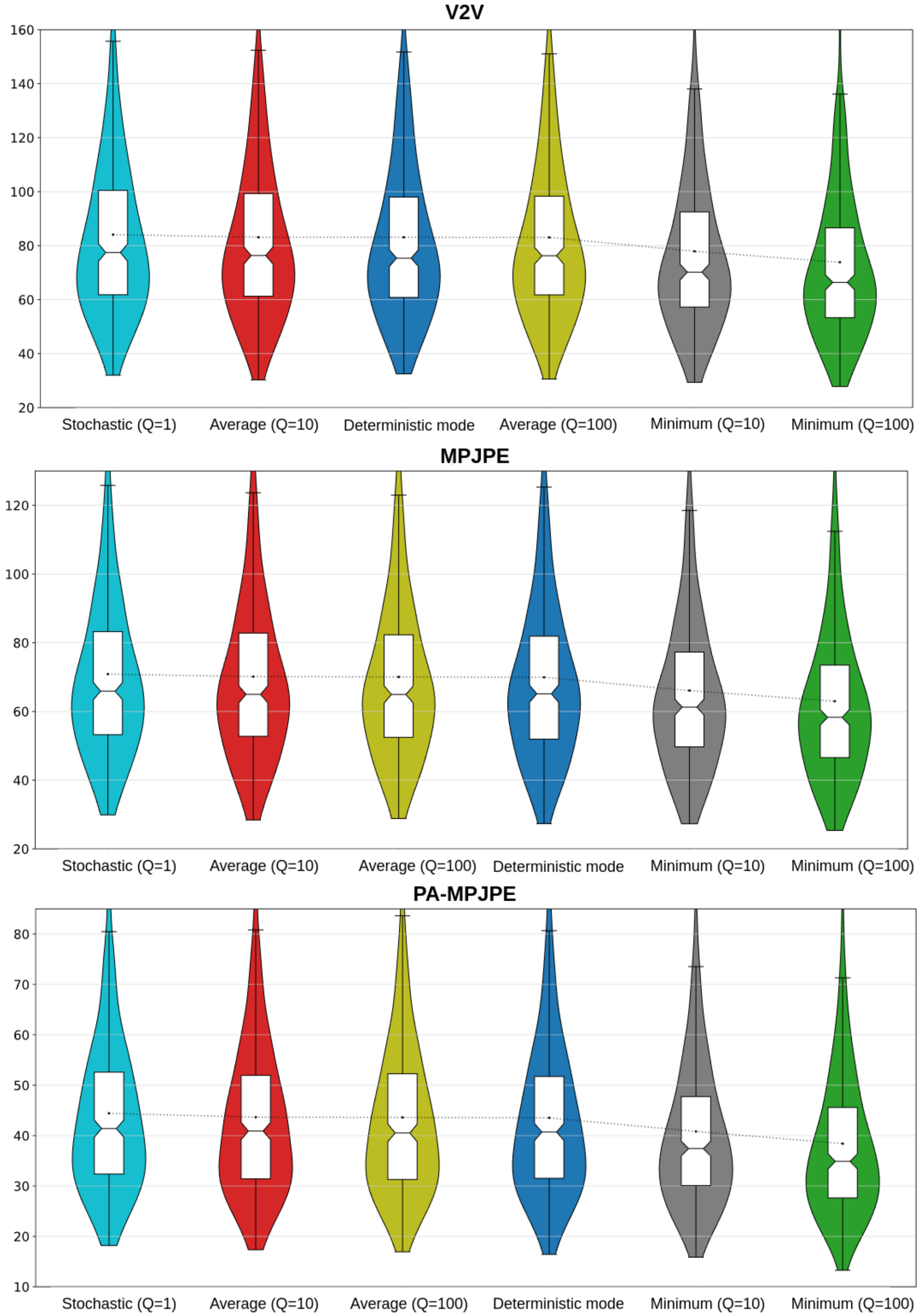

Figure 6: **Error distribution.** We visualize the distribution of the error on 3DPW.

the importance of having a deterministic mode for rapid and accurate predictions, which can be considered an estimator of the average prediction over Q samples.

When selecting the prediction with the minimum error among N samples, we observe a shift in the distribution shapes, with errors concentrated toward lower values. While the highest error values decrease notably, the lowest remain relatively unchanged. This phenomenon likely occurs because the lowest errors typically correspond to meshes that are easier to predict and exhibit lower uncertainty. Consequently, the stochastic mode proves particularly beneficial for challenging images, where uncertainty is higher and multiple predictions offer valuable insights.

C Experimental details

Pre-training stage. As detailed in the main body, the pre-training stage is done on a subset of AMASS [Mahmood et al., 2019], as introduced in [Black et al., 2023]. We pre-train MEGA for 500 epochs, a task accomplished in less than a day on 4 A100 GPUs. MEGA is trained using the AdamW optimizer [Loshchilov and Hutter, 2017] with a cosine scheduler to adjust the learning rate. The base learning rate is $1e-3$, and we have a warmup of 20 epochs. The optimizer's parameters are $\beta_1 = 0.9$, $\beta_2 = 0.99$, and the weight decay is 0.05.

Supervised training stage. The supervised training stage is done on a mix of standard datasets for HMR as presented in Sec. 4.1. We first train MEGA on MSCOCO [Lin et al., 2014] for 100 epochs and then train on the whole training set for 10 epochs. Each step takes about 1 day on 4 A100 GPUs. The training settings are exactly the same as the pre-training regarding learning rates and schedulers. For each training step, we start at epoch 0. Note that we have a lower learning rate in practice for the training on the mix of datasets because we stop the training before finishing the warming steps. For training with HRNet and ResNet-50 backbones, the weights of the backbone are fine-tuned with the same settings as the other parameters of MEGA. When using ViT, the backbone is frozen during the training on MSCOCO, and we only fine-tune the last 10 blocks when training on standard datasets for computing power reasons.

HMR. For recovering human meshes from images in deterministic mode, we predict all images in a single step without randomness. In stochastic mode, we have to set the number of steps for generating the sequence of human mesh tokens and the amount of noise injected for the Gumble-max sampling. Note that we did not test MEGA with a ViT in stochastic mode. With HRNet, we generate meshes in 5 steps, and the initial noise temperature is 1. The generation process with ResNet-50 is made in 2 steps with an initial noise 10. The amount of noise at step t is $A \times (1 - \frac{t}{T})$ where A is the initial noise temperature.

Random meshes generation. We generate random meshes in 20 steps. We want the generation to be completely random for the first steps so that the predictions are diverse. However, the last steps should be almost deterministic to obtain realistic meshes. The initial noise temperature is $A = 1.2$, and the amount of noise at step t is given by $(A \times (1 - \frac{t}{T}))^6$.

D Generation process

We provide visualizations of the prediction process in stochastic mode and for random mesh generation. Specifically, for the HMR task, using the model with an HRNet backbone that generates meshes in 5 steps, given an image, we visualize all intermediate steps. In the case of random generation, we illustrate the meshes after 1, 5, 9, 13, 16, and 20 steps. At each step, all masked tokens are set to 0. Results are shown in Fig. 7. Importantly, our model does not predict unrealistic meshes in the first steps, as all masked tokens were replaced by the index 0.

For the HMR visualization, we observe that a rough estimate of the mesh is provided even in the initial steps, where only a few indices are predicted (3 during the first iteration and 7 during the second). This outcome is expected because the tokens set to be visible after each step are the most likely. Consequently, the indices with the highest certainty are selected first, enabling the construction of a preliminary mesh estimate. Subsequent steps refine these predictions, enhancing their realism.

In random mesh generation, the initial mesh appears almost identical across all generations, as only one token is predicted in the first step (with all others set to 0 for visualization). However, diversity quickly emerges in subsequent steps, with the final steps refining the meshes to be more realistic.

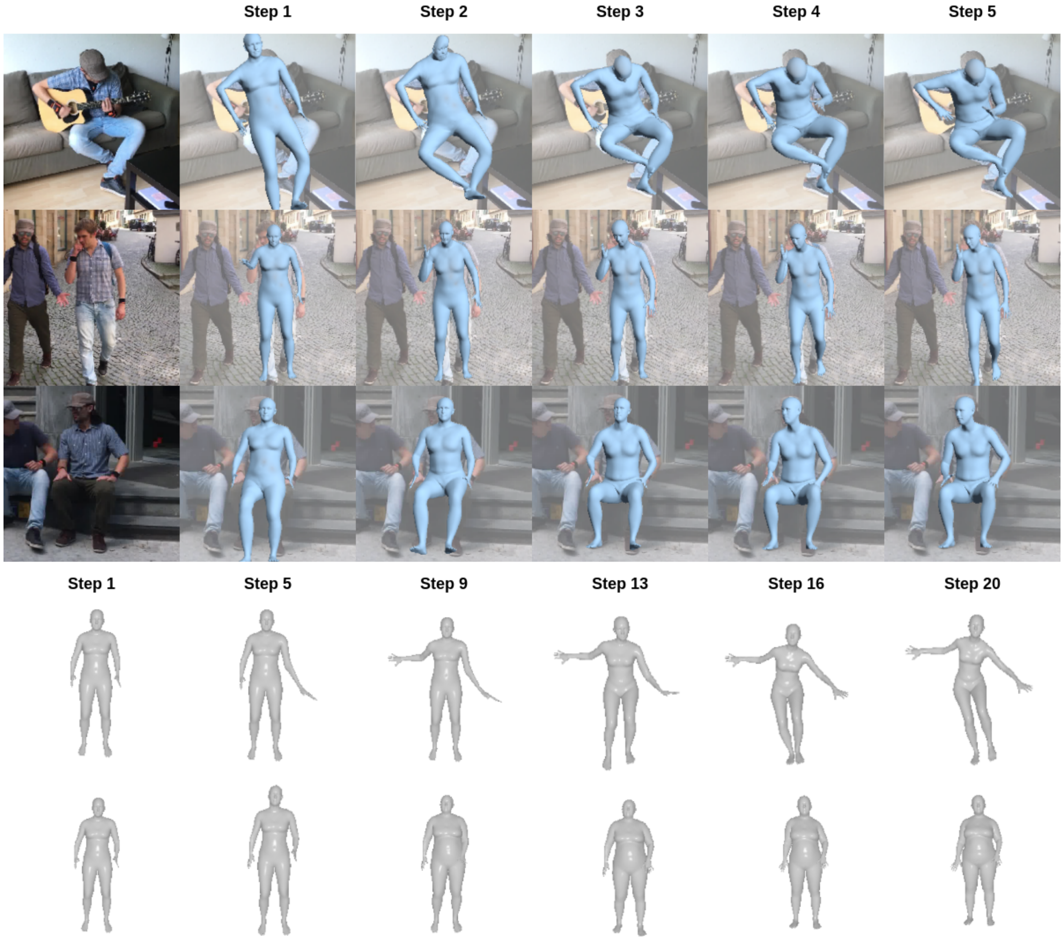

Figure 7: **Prediction process iterations.** We visualize the predictions for intermediate steps in stochastic mode (first 3 rows) and intermediate random generations (last 2 rows). All masked tokens are replaced by the first token of the codebook, corresponding to index 0.

This pattern was anticipated, as the initial steps involve considerable randomness, whereas the later steps tend to become more deterministic.

E Qualitative results and failure cases

We present several failure cases in Fig. 8. Extreme poses can result in prediction errors, occasionally leading to non-anthropomorphic predictions due to the non-parametric nature of our approach. Notably, uncertainty is particularly high in such instances.

Fig. 9 presents qualitative samples from in-the-wild datasets. We can observe that in some cases (for instance, images in the second row and left image in the fourth row), our predictions appear even more accurate than the ground truth. While this result is encouraging, it underscores the limited value of striving for fractions of millimeters of accuracy on datasets like 3DPW when the ground truth itself is imperfect. Additionally, we observe that uncertainty is significantly higher for occluded body parts, which often correspond to body extremities.

Supplementary References

[Black et al., 2023] Black, M. J., Patel, P., Tesch, J., and Yang, J. (2023). BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737.

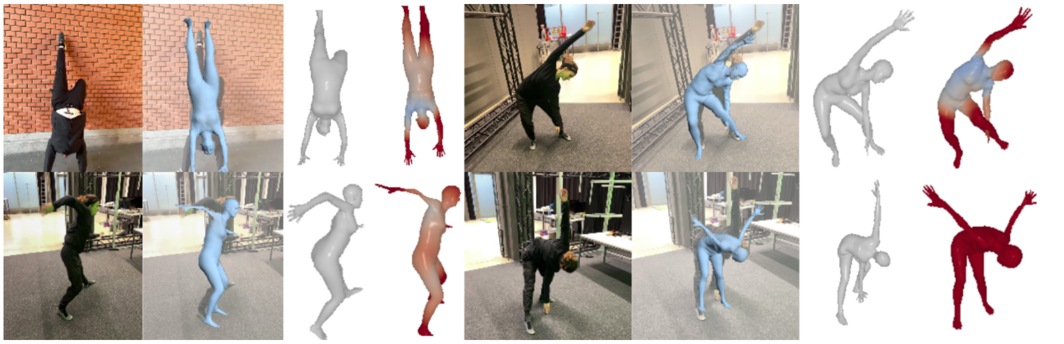

Figure 8: **Failure cases.** In failure cases, it is worth noting that our model predicts high uncertainty.

[Choi et al., 2022] Choi, H., Moon, G., Park, J., and Lee, K. M. (2022). Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1484.

[Kocabas et al., 2021] Kocabas, M., Huang, C.-H. P., Hilliges, O., and Black, M. J. (2021). PARE: Part attention regressor for 3D human body estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137.

[Kolotouros et al., 2019] Kolotouros, N., Pavlakos, G., Black, M. J., and Daniilidis, K. (2019). Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Mahmood et al., 2019] Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5442–5451.

[Sun et al., 2021] Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M. J., and Mei, T. (2021). Monocular, one-stage, regression of multiple 3D people. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11179–11188.

[von Marcard et al., 2018] von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., and Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 601–617. Springer.

[Yang et al., 2023] Yang, C., Kong, K., Min, S., Wee, D., Jang, H.-D., Cha, G., and Kang, S. (2023). SEFD: learning to distill complex pose and occlusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14941–14952.

[Yao et al., 2022] Yao, C.-H., Yang, J., Ceylan, D., Zhou, Y., Zhou, Y., and Yang, M.-H. (2022). Learning visibility for robust dense human body estimation. In *European Conference on Computer Vision (ECCV)*, pages 412–428. Springer.

[Zhang et al., 2020] Zhang, T., Huang, B., and Wang, Y. (2020). Object-occluded human shape and pose estimation from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7376–7385.

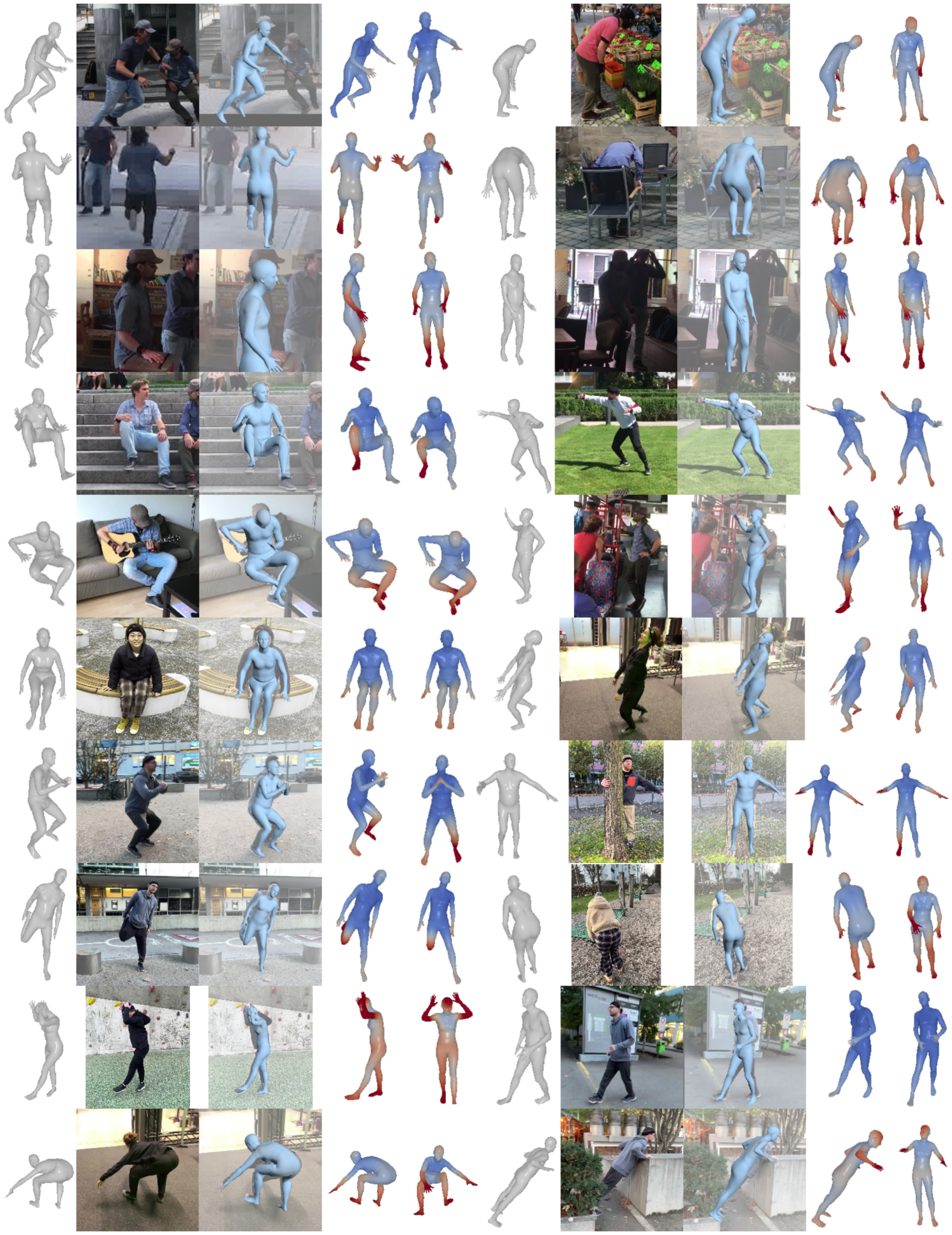

Figure 9: **Qualitative examples on 3DPW and EMDB.** The first column is the groundtruth, the second and third are the image and the reprojection of the deterministic projection, and the third and fourth are uncertainty computed in stochastic mode and visualized in oriented and canonical points of view.