VQ-HPS: Human Pose and Shape Estimation in a Vector-Quantized Latent Space

Guénolé Fiche[1], Simon Leglaive[1], Xavier Alameda-Pineda[2], Antonio Agudo[3], and Francesc Moreno-Noguer[3]

[1]CentraleSupélec, IETR UMR CNRS 6164, France
[2]Inria, Univ. Grenoble Alpes, CNRS, LJK, France
[3]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Spain

Abstract

Previous works on Human Pose and Shape Estimation (HPSE) from RGB images can be broadly categorized into two main groups: parametric and non-parametric approaches. Parametric techniques leverage a low-dimensional statistical body model for realistic results, whereas recent non-parametric methods achieve higher precision by directly regressing the 3D coordinates of the human body mesh. This work introduces a novel paradigm to address the HPSE problem, involving a low-dimensional discrete latent representation of the human mesh and framing HPSE as a classification task. Instead of predicting body model parameters or 3D vertex coordinates, we focus on predicting the proposed discrete latent representation, which can be decoded into a registered human mesh. This innovative paradigm offers two key advantages. Firstly, predicting a low-dimensional discrete representation confines our predictions to the space of anthropomorphic poses and shapes even when little training data is available. Secondly, by framing the problem as a classification task, we can harness the discriminative power inherent in neural networks. The proposed model, VQ-HPS, predicts the discrete latent representation of the mesh. The experimental results demonstrate that VQ-HPS outperforms the current state-of-the-art non-parametric approaches while yielding results as realistic as those produced by parametric methods when trained with few data. VQ-HPS also shows promising results when training on large-scale datasets, highlighting the significant potential of the classification approach for HPSE. See the project page at https://g-fiche.github.io/research-pages/vqhps/.

1 Introduction

Capturing and understanding human motion from RGB data is a fundamental task in computer vision, with many applications such as character animation for the movie and video-game industries [4, 5, 6] or performance optimization in sports [7, 8]. However, due to depth ambiguity, estimating 3D human pose and shape from monocular images is an underdetermined problem. To overcome this issue, parametric approaches (also called model-based) use statistical models of the human body, which enable the reconstruction of a 3D human mesh by predicting a small number of parameters [9, 10, 11, 12, 13]. Earlier methods were optimization-based, estimating the parameters of a human body model iteratively using 2D cues [14, 15, 16]. However, their need for a good initialization, slow running time, and propensity to converge towards local minima led many recent works to focus on regression-based methods, which predict the parameters of a human body model directly from RGB data [1, 2, 17]. Despite producing realistic results in most scenarios, methods regressing the parameters of a human body model face several issues well documented in the literature: 1) Parametric methods struggle in capturing detailed body shape and are biased towards the mean shape [18]; 2) Most human body models use rotations along the kinematic tree for expressing the pose. In addition to being difficult to predict for neural networks [19, 20], this representation induces error accumulation when all rotations are predicted simultaneously [21, 22]; 3) Most regression methods extract global feature vectors from the image as an input, which do not contain fine-grained local details [23].

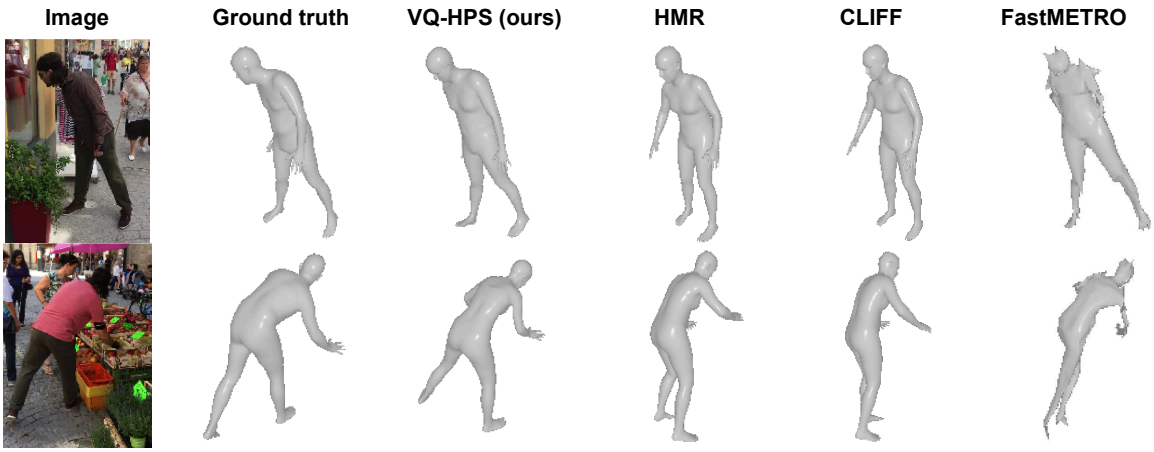

Figure 1: **VQ-HPS formulates the human pose and shape estimation problem as a classification task in a vector-quantized latent space**. We present the results of VQ-HPS on two challenging scenarios with in-the-wild conditions and poor illumination, comparing its performance to that of HMR [1], CLIFF [2] and FastMETRO-S [3] when trained on little data.

To alleviate these issues, several works switched to methods inspired by 3D pose estimation models that predict 3D coordinates directly. Earlier methods predicted the 6890 vertices of the full SMPL [9] mesh using graph convolutional neural networks (GCNNs) modeling the mesh structure and focusing on local interaction between neighboring vertices [20, 19]. While [24] used Transformers [25] to model global interactions between joints and vertices, others argued that a hybrid architecture mixing Graph Convolutional Neural Networks (GCNNs) and Transformers would enable modeling both local and global interactions [23]. More recently, FastMETRO [3] proposed a Transformer-based encoder-decoder architecture to disentangle image and mesh features and to predict 3D coordinates of body joints and a coarse mesh that can be upsampled to the full SMPL body mesh. Significantly different from prior works, LVD [18] proposed an optimization-based approach estimating each vertex position independently by predicting vertex displacement with neural fields. Despite proposing alternatives to model-based approaches, these methods also present some drawbacks: 1) Approaches regressing all vertices of the body mesh at once lack global interaction modeling when using GCNNs [23] and have a very high computational cost when using Transformers [3, 26]; 2) Regression-based methods sometimes output noisy meshes, some of them regress the SMPL parameters from the predicted mesh to obtain smoother predictions, but it comes with a loss of accuracy [3, 19, 20]. This problem is even more glaring when little training data is available, with non-anthropomorphic predictions as displayed in Fig. 1 for[3]; 3) Methods regressing 3D vertices are very sensible to the distribution shift between training and test data [24] (see Sec. 5.4); 4) LVD [18] is real-time and obtains state-of-the-art results for shape estimation, but is not adapted to extreme poses.

This work introduces a method significantly different from all prior human pose and shape estimation (HPSE) approaches. Instead of predicting the parameters of a human body model or 3D coordinates, we learn to predict a discrete latent representation of 3D meshes, transforming the HPSE into a classification problem in which we can exploit the originally targeted discriminative power of Transformers, which has been proven unmatched in natural language processing. For learning our discrete latent representation of meshes, we build on the vector quantized-variational autoencoder (VQ-VAE) [27] framework and adapt it to the fully convolutional mesh autoencoder proposed in [28]. The encoder of the proposed model, called Mesh-VQ-VAE, provides a low-dimensional discrete latent representation preserving the spatial structure of the mesh. We then propose a Transformer-based encoder-decoder model, called VQ-HPS, for learning to solve the HPSE problem using the cross-entropy loss. Once the mesh discrete representation is predicted, we can decode it using the pre-trained Mesh-VQ-VAE decoder and obtain a full mesh following the SMPL mesh topology [9]. Since the Mesh-VQ-VAE is pre-trained on a large human motion capture database [29], it automatically learns to decode smooth and realistic human meshes. This is particularly interesting when training with little data: VQ-HPS learns to predict sequences of indices corresponding to realistic meshes early in the training process, as demonstrated in the supplementary materials.

In the context of few training data availability, VQ-HPS achieves state-of-the-art performance on the challenging in-the-wild 3DPW [30] and EMDB [31] benchmarks: it significantly outperforms other methods quantitatively while producing qualitative results as realistic as parametric methods (see Fig. 1). Moreover, it also shows SOTA results when trained on standard large-scale datasets, enhancing the significant potential of the classification-based approach for solving the HPSE problem.

Our key contributions can be summarized as follows:

- A Mesh-VQ-VAE architecture providing a discrete latent representation of 3D meshes.
- A classification-based formulation of the HPSE problem using the introduced discrete latent representation of human meshes.
- VQ-HPS, a Transformer-based encoder-decoder model learning to solve the proposed HPSE classification problem using the cross-entropy loss.
- Code and trained models are available from the project page.

2 Related Work

2.1 Parametric Approaches

Several methods are dedicated to recovering the parameters of a parametric human model, such as SMPL [9]. Optimization techniques iteratively estimate the parameters of a body model based on images or videos, ensuring that the projection of predictions aligns with a set of 2D cues, including 2D skeletons [12, 14, 32, 33], part segmentation [34, 15], or DensePose [35]. Pose and motion priors are commonly incorporated into optimization methods to enhance the realism of predictions [16, 36, 37, 38]. On the contrary, regression methods employ neural networks to predict the parameters of a human body model from input images or videos. Many of these methods leverage convolutional neural networks (CNNs) for extracting image features [39, 40, 1, 41, 42, 2, 43, 44, 45, 46, 47, 48]. Recent works have demonstrated remarkable performance by replacing CNNs with Vision Transformers [49] as seen in [17, 50, 51, 52]. Some methods output probabilistic results, enabling sampling among plausible solutions [53, 54, 55, 56, 57]. While optimization methods typically yield superior results, they come with significantly longer running times than regression methods and require precise initialization and accurate 2D cues. One limitation in training regression models is the scarcity of RGB data with 3D annotations. Prior works have addressed this challenge by employing synthetic data [58, 59, 60, 53, 61] or pseudo-labels [33, 62, 15] for training their models.

While parametric models can estimate reasonable human poses, the model parameter space may not be the most suitable focus for predicting human pose and shape [19, 18]. Recognizing these limitations inherent in parametric approaches has spurred the development of non-parametric methods.

2.2 Non-parametric Approaches

Several works have explored methods for directly predicting 3D meshes without relying on the parameters of a human body model [19, 18, 63, 23, 24, 3]. In earlier approaches, regression architectures based on GCNNs were proposed, utilizing a graph structure derived from the topology of the SMPL human mesh [19, 63, 23]. Recent advancements have leveraged Transformer architectures, capitalizing on attention mechanisms to capture relationships between joints and vertices. While approaches like [23, 24] have introduced encoder-based strategies that concatenate image features and mesh tokens for predicting 3D coordinates, FastMETRO [3] presented an encoder-decoder architecture, effectively disentangling image and mesh modalities. Recently, [26] introduced a token pruning strategy to enhance the efficiency of Transformer-based HPSE, and [18] achieved state-of-the-art accuracy in body shape estimation through an optimization-based approach relying on per-vertex neural features.

This work introduces a non-parametric approach to HPSE. Our objective is to estimate the vertices of a human body mesh, adhering to the SMPL topology [9]. In contrast to all prior works, our method involves predicting the mesh through a discrete latent representation, reframing HPSE as a classification problem. Although exploiting the discriminative power of classification networks has already been proposed for the Human Pose Estimation (HPE) problem (see Sec. 2.3), to our knowledge, this has not been done before for the HPSE problem.

2.3 Quantization of the Human Pose and Shape

Some works explored quantization for HPE. [64] proposed to discretize horizontal and vertical coordinates for 2D HPE. On the other hand, [65, 66] used anchor poses and refined them for solving the 3D HPE problem. [67] proposed a human pose and shape classification method, but the system was trained on only 12 different postures. Some approaches proposed hand shape classification [68], especially for sign-language recognition following the works of [69]. Some works also proposed face shape classification [70] and head pose estimation [71] using a Support Vector Machine.

Recent works in human motion generation [72, 73, 74, 75] used a VQ-VAE [27] for quantizing human motion. The main difference with the proposed Mesh-VQ-VAE is that a single index encodes a sequence of poses in these works.

In contrast, we use several indices to encode a single pose, allowing for higher precision. Also, none of these works encode the 3D mesh: [72] uses the SMPL parameters, and others only encode a 3D skeleton. Furthermore, human motion forecasting and generation tasks differ significantly from HPSE.

Concurrently with the present work, TokenHMR [76] proposes to quantize human pose for HPSE. As in our case, this tokenization acts as a pose prior, using a dictionary of valid pose tokens. However, the tokenization of TokenHMR differs from ours as it happens on the SMPL [9] pose parameter while we quantize the 3D mesh. Another major difference between VQ-HPS and TokenHMR is how the tokenized pose is used. TokenHMR uses it as an intermediate representation while still solving the HPSE problem as a regression task: similar to prior works in HPSE, the training targets are the SMPL pose and shape parameters and the vertices' coordinates. We propose to frame HPSE as a classification task: the unique training target for the mesh is its discrete quantized representation. Experiments show that VQ-HPS achieves better results than TokenHMR using less training data and a less powerful backbone.

3 Background

SMPL model. SMPL [9] is a skinned vertex-based human body model that maps the body shape parameter $\beta \in \mathbb{R}^{10}$ and the pose parameter $\theta \in \mathbb{R}^{72}$ to 3D vertices through the differentiable function $\mathcal{M}(\beta, \theta)$. It outputs the 3D vertices $V \in \mathbb{R}^{6890\times3}$ of a registered mesh, and 3D joints $J \in \mathbb{R}^{24\times3}$ can be extracted from the mesh using the joint regressor matrix $\mathcal{J}_{smpl}$. In this work, we do not predict the parameters of the SMPL model, as prior works [19, 20, 18] showed that they are not a suitable target for regression models. However, the mesh predicted by the proposed VQ-HPS model follows the SMPL mesh topology. It allows us to use tools like joint regressors and provides a fair comparison with existing approaches.

Fully convolutional mesh autoencoder. The fully convolutional mesh autoencoder [28] is an autoencoder specifically tailored for handling arbitrary registered mesh data. It relies on the definition of novel convolution and pooling operators with globally shared weights and locally varying coefficients depending on the mesh structure. These variable coefficients are pivotal in capturing intricate details inherent to irregular mesh connections, contributing to the model's performance in mesh reconstruction. One of the main advantages of fully convolutional architecture is that the latent codes are localized, which gives a latent space preserving the spatial structure of the mesh. The latent representation of the fully convolutional mesh autoencoder lies in $\mathbb{R}^{N\times L}$ where N is the number of latent vectors, and L is the dimension of latent vectors.

Vector quantized-variational autoencoder. The VQ-VAE [27] is an encoder-decoder model with a discretized latent space. The idea is to learn jointly an encoder, a dictionary of latent codes, and a decoder. The encoder maps the input data x into a latent variable $z \in \mathbb{R}^{N\times L}$. We then discretize z using a learned dictionary of S latent codes of dimension L. We can then write $z_d \in \mathbb{R}^{N\times L}$ where each vector of z is replaced by the closest latent code, or $z_q \in \{1, \dots, S\}^N$, where the index of the closest latent code replaces each vector of z. The decoder reconstructs x from the discrete latent representation z_d and the learned dictionary.

4 Method

4.1 Proposed HPSE method

We propose a novel classification-based method for HPSE. Our goal is to predict an oriented 3D mesh from an image. VQ-HPS consists of an encoder-decoder architecture, predicting the human mesh discrete representation of the introduced Mesh-VQ-VAE from image features. We believe this is the most adapted architecture for predicting our discrete latent representation, with encoder tokens corresponding to image patches and decoder tokens corresponding to body parts and indices in the latent space. To ease the low-dimensional representation learning of the mesh, the predicted mesh is non-oriented and centered on the origin (see Appendix A): we call it a *canonical mesh*. To obtain the final oriented mesh, we then need to predict the rotation $R \in \mathbb{R}^{3\times3}$, and for better alignment with the image, we also regress the perspective camera $\pi = [s, t] \in \mathbb{R}^3$ where s is a scale parameter and t is a 2D translation. The overall method is shown in Fig. 2, and we will now detail each of its primary components.

Mesh-VQ-VAE. For learning discrete representations of meshes, we build on the fully convolutional mesh autoencoder [28] (see Sec. 3) for encoding the full canonical mesh vertices $V_c \in \mathbb{R}^{6890\times3}$ to a latent representation $z \in \mathbb{R}^{N\times L}$. We add a vector quantization step in the latent space similar to [27] (see Sec. 3), which maps z to the discrete latent representation $z_q \in \{1, \dots, S\}^N$. While the fully convolutional architecture preserves the spatial structure of the mesh, the added quantization step allows us to view the HPSE as a classification task as we aim to predict the indices of

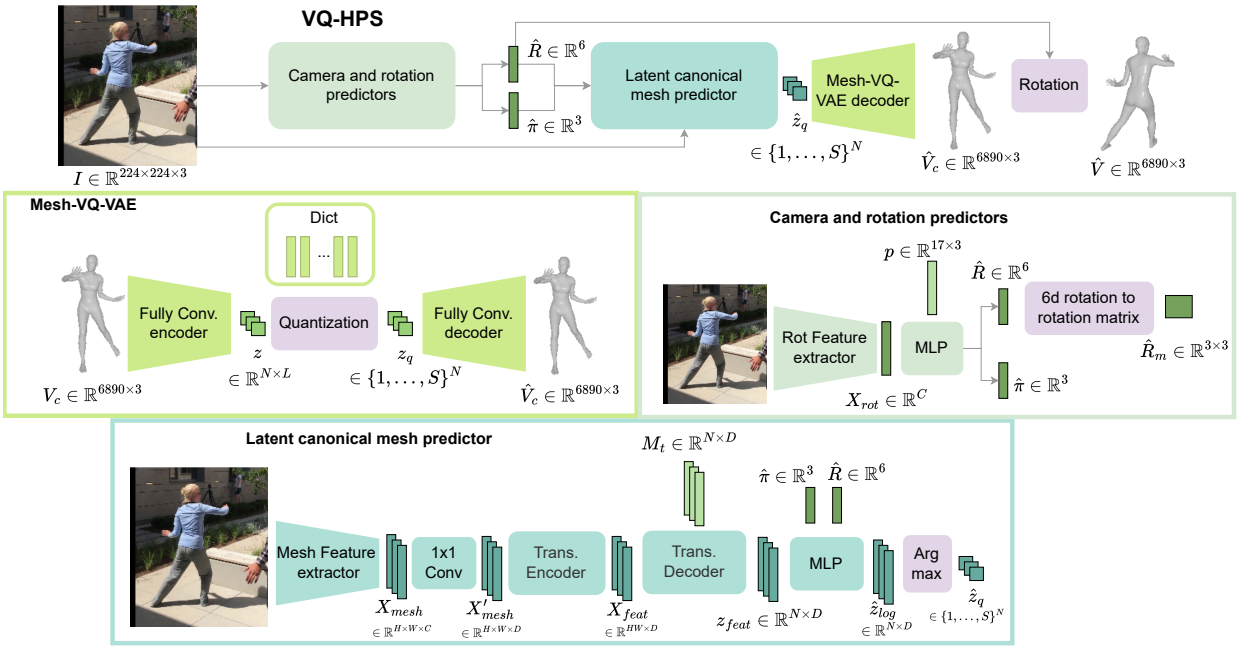

Figure 2: **VQ-HPS global process for predicting the mesh given an image.** We first predict the camera $\hat{\pi}$ and the rotation $\hat{R}$ from the image I. Then, we use the image, the predicted rotation, and the camera to predict the vertices $\hat{V}_c$ of the canonical mesh. Finally, $\hat{V}_c$ is rotated according to $\hat{R}$ to obtain the final mesh vertices $\hat{V}$.

the latent mesh representation given an image. Our Mesh-VQ-VAE in Fig. 2 can be seen as a VQ-VAE [27] whose architecture corresponds to the fully convolutional mesh autoencoder.

Feature extractors. The first step for image-based HPSE is to extract features from the image. We use CNN backbones to preserve the spatial structure of the image, and we obtain features $X \in \mathbb{R}^{H \times W \times C}$, where C is the number of channels of the backbone and H and W are the spatial dimension. We use two feature extractors. The feature extractor of the camera and rotation predictors gives X_rot. The feature extractor gives X_{mesh} with $W = 1$ in the latent canonical mesh predictor.

Rotation and camera prediction. We start by predicting the mesh rotation and the perspective camera parameters (see again Fig. 2). These predictions depend on the image features and an initial body pose $p \in \mathbb{R}^{17 \times 3}$ following the Human3.6M [77] joints layout and corresponding to the SMPL T-pose. We predict the rotation $\hat{R}$ and the weak perspective camera parameters $\hat{\pi}$.

Latent canonical mesh regressor encoder. We then predict the discrete latent representation of the canonical mesh. The Transformer encoder inputs are the features extracted by the CNN backbone. Before being fed to the Transformer encoder, we apply a 1x1 convolution on the image features to make them of dimension and obtain $X'_{mesh} \in \mathbb{R}^{H \times W \times D}$ where D is the hidden state size of the Transformer. These features are flattened to obtain HW tokens of dimension D, and then we add positional encoding. The obtained tokens are fed to a Transformer encoder, using self-attention between all image tokens to output encoded image features $X_{feat} \in \mathbb{R}^{HW \times D}$.

Latent canonical mesh regressor decoder. The Transformer decoder takes as inputs N learned mesh tokens M_t of size D, each responsible for predicting an index of the Mesh-VQ-VAE discrete latent representation. We need to solve N classification problems, one for each index. Each problem has S classes, S corresponding to the size of the Mesh-VQ-VAE codebook. The Transformer decoder consists of self-attention between learned tokens and cross-attention with image features. It outputs latent mesh features $z_{feat} \in \mathbb{R}^{N \times D}$. Then (see Fig. 2), to obtain the logits $\hat{z}_{log} \in \mathbb{R}^{N \times S}$, we rely on the mesh features as well as on the previously predicted rotation and camera. We obtain the predicted discrete representation $\hat{z}_q \in \{1, \dots, S\}^N$ by applying an $\arg\max(\cdot)$ operation.

Reconstructing the full mesh. From the discrete latent mesh representation $\hat{z}_q \in \{1, \dots, S\}^N$, we use the decoder of the introduced Mesh-VQ-VAE to reconstruct the vertices of a full canonical mesh $\hat{V}_c \in \mathbb{R}^{6890 \times 3}$. We apply the

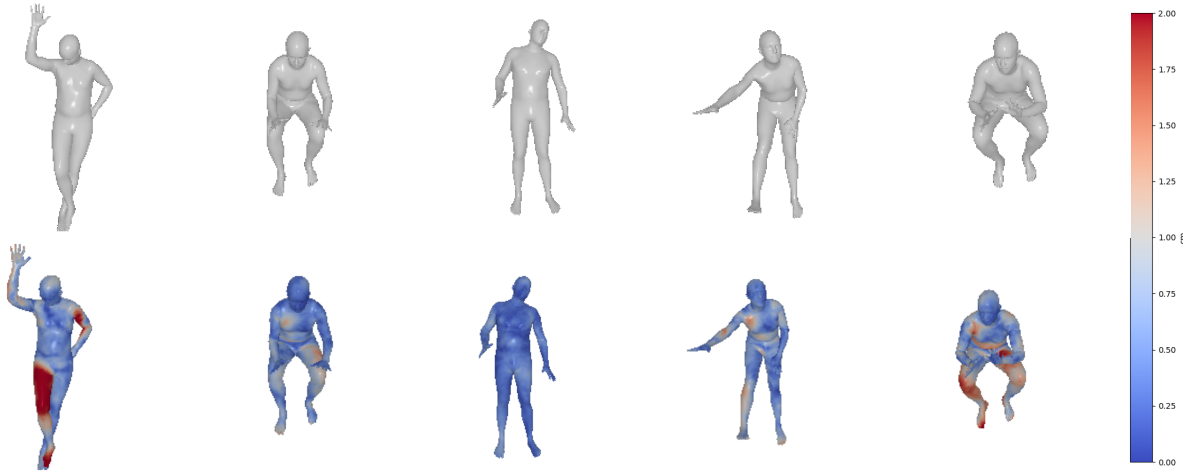

Figure 3: **Mesh-VQ-VAE reconstruction error.** Samples of reconstruction on the 3DPW test set. The error is in cm and corresponds to the Euclidean distance between the reconstruction's original mesh and the corresponding vertex.

predicted rotation $\hat{R}$ of the oriented mesh in the frame coordinates to the vertices to obtain the vertices $\hat{V}$. This process is shown in Fig. 2.

4.2 Training VQ-HPS

VQ-HPS is trained in a supervised manner, given a dataset of RGB images paired with meshes. The canonical mesh predictor is trained solely on the discrete latent representation of meshes. To obtain the latent representation of the ground truth and decode the predicted indices to a full mesh, we use the Mesh-VQ-VAE, which is pre-trained and frozen during the VQ-HPS training. The fact that the Mesh-VQ-VAE is frozen during the training of VQ-HPS is crucial for making realistic predictions in the context of scarce data. Pre-training acts as a regularization, allowing for the reduction of the amount of training data.

Mesh-VQ-VAE. The Mesh-VQ-VAE (see Fig. 2) is trained on the AMASS [29] dataset and finetuned on the 3DPW [30] training set. To ease the learning of the mesh discrete representation with a limited number of indices, we train the Mesh-VQ-VAE with non-oriented meshes translated to the origin (canonical meshes). The final reconstruction error is 4.7 mm. This reconstruction error is an important parameter as it corresponds to the minimal per-vertex error (see Sec. 5.2) we can obtain. Qualitative reconstruction results on 3DPW are shown in Fig. 3.

Latent canonical meshes. For learning to predict the pose and shape, we only use the discrete representation of the canonical mesh as the training target. The loss $\mathcal{L}_{mesh}$ is the cross-entropy between the discrete latent representation of the ground truth mesh and the prediction.

Mesh rotation. We learn to predict the global orientation by computing the mean squared error between the ground truth and predicted rotation matrices.

Reprojection. We add a reprojection error to guide the rotation learning and for better image alignment. It is computed between the 2D projection (using the predicted weak-perspective camera) of the 3D joints extracted from the predicted mesh and the 2D ground truth joints. This loss is computed using the SMPL 24 joints, which can be extracted from the full mesh using a joint regressor $\mathcal{J}_{smpl}$ (see Sec. 3). The reprojection loss is computed as:

$$\mathcal{L}_{2D} = ||\hat{s}\Pi(\hat{J}_{3D}) + \hat{t} - J_{2D}||_1, \tag{1}$$

where $\hat{s}$ is the predicted scale, $\hat{t}$ the predicted 2D translation and $\hat{J}_{3D}$ are the 3D joints computed from the predicted oriented mesh vertices $\hat{V}$. Π is the orthographic projection using the matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^\top$ and J_{2D} denotes the ground truth 2D joints.

Learning scheme. The rotation prediction is learned using $\mathcal{L}_{rot}$ and $\mathcal{L}_{2D}$. We use $\mathcal{L}_{2D}$ for the camera and $\mathcal{L}_{mesh}$ for the canonical mesh. $\mathcal{L}_{2D}$ might help to learn the pose, but we chose not to use it to demonstrate that the cross-entropy is sufficient for making accurate predictions.

Table 1: **Results on in-the-wild datasets** We compare VQ-HPS with SOTA methods using standard metrics on 3DPW trained with 3DPW (1st col.), 3DPW trained with COCO (2nd col.), EMDB trained with EMDB (3rd col.), and EMDB trained with COCO (4th col.). All results are given in mm.

Method	PVE ↓				MPJPE ↓				PA-MPJPE ↓			
HMR [1]	209.3	110.3	191.3	170.0	177.6	94.1	172.2	149.2	89.3	57.8	96.2	82.4
CLIFF [2]	223.9	105.4	163.3	163.2	188.8	89.3	144.7	143.0	89.2	56.8	86.9	81.6
FastMETRO-S [3]	176.3	107.8	151.0	**143.1**	157.0	95.8	132.9	**123.9**	104.6	57.0	95.5	80.2
VQ-HPS (ours)	**163.9**	**102.9**	**138.5**	152.7	**139.8**	**88.0**	**117.1**	131.1	**84.9**	**53.3**	**77.5**	**74.5**

5 Results

5.1 Datasets

AMASS. The Mesh-VQ-VAE is trained on AMASS [29], a large human motion database in the SMPL [9] format. It contains more than 11000 motions and 300 subjects, which makes it representative of the variety of body poses and shapes.

3DPW. This dataset [30] consists of 60 in-the-wild RGB videos with 3D ground truth for human bodies. We use the pre-defined splits for training, validation, and testing. Note that when training on a mix of datasets (see Sec. 5.4), we do not finetune models on 3DPW to assess generalization. We also evaluate VQ-HPS on 3DPW-OCC, a benchmark proposed in [78] containing videos of 3DPW with occlusions.

EMDB. EMDB [31] contains 81 indoor and outdoor videos with the ground truth SMPL bodies. We use EMDB1, which consists of the 17 most challenging sequences for testing, and train on the rest of the dataset (referred to as EMDB2 in [31]).

COCO. COCO [79] is a dataset of images annotated with 2D keypoints. For training a human mesh predictor, we follow [33, 2] and use pseudo-ground truth meshes. We use the same annotations as [2], with 28k images.

5.2 Metrics

We use several metrics to evaluate the predictions of VQ-HPS. All of them will be expressed in millimeters (mm) for the whole results section.

Per-vertex error (PVE) measures the Euclidean distance between the predicted vertices and the ground truth.

Mean-per-joint error (MPJPE) measures the Euclidean distance between the predicted joints and the ground truth. In our case, the joints are extracted from the predicted mesh using a joint regressor similar to $\mathcal{J}_{smpl}$.

Procrustes-aligned mean-per-joint error (PA-MPJPE) measures the Euclidean distance between the predicted joints and the ground truth after a Procrustes alignment.

5.3 Training on limited data

We train VQ-HPS separately on the 3DPW, COCO, and EMDB training sets (see Sec. 5.1) to see how it performs when trained on limited data. We compare our performance with HMR [1], CLIFF [2], and FastMETRO [3] trained with the same data. We chose these 3 models for comparison because HMR is the basic architecture for parametric human mesh recovery, CLIFF is the SOTA for parametric HPSE, and FastMETRO is the SOTA for non-parametric HPSE and the closest method to ours. For these experiments, the backbone for all networks is ResNet-50 [80] pre-trained on ImageNet [81]. We use the public implementation of FastMETRO and adapt the HMR and CLIFF implementations provided by [60]. For all comparisons, we use FastMETRO-S, as this version is the closest to VQ-HPS. When training on COCO, we propose an improved version of VQ-HPS, replacing the MLP of the latent canonical mesh regressor with a Transformer implementing self-attention between the latent mesh features z_{feat}. Quantitative results are shown in Tab. 1, and visualizations are available in Fig. 4.

Overall, VQ-HPS outperforms the SOTA methods significantly when training on 3DPW and EMDB (see Tab. 1, col. 1 and 3). Visualization of the results confirms that our method performs best (see Fig. 4), and we propose a more detailed analysis of the error in Appendix D. HMR and CLIFF show realistic predictions but are less accurate than VQ-HPS. Despite rather good metrics, FastMETRO produces non-smooth results that do not correspond to human body shapes.

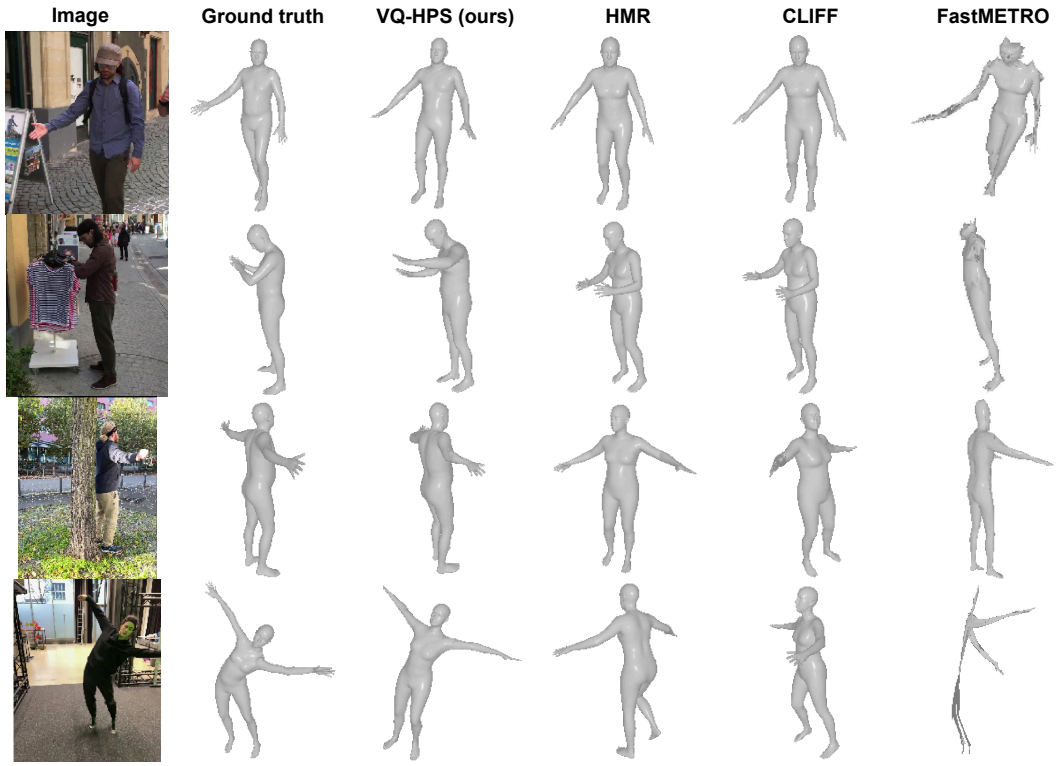

Figure 4: **Qualitative results** We compare our method with HMR [1], CLIFF [2] and FastMETRO-S [3] on 3DPW trained on 3DPW (first 3 rows), and EMDB trained on EMDB.

This is probably due to the limited training sets, as the results displayed in the original paper looked more realistic. This highlights a clear limitation of non-parametric approaches predicting the 3D coordinates; in comparison, VQ-HPS needs much less data to provide realistic predictions, probably because learning sequences of indices corresponding to anthropomorphic meshes is easier than understanding the structure of the 3D vertices. Furthermore, VQ-HPS benefits from the large-scale pre-training on human motion datasets, acting as a regularization to allow learning with fewer image data labeled with 3D poses.

The fact that the Mesh-VQ-VAE, whose decoder is an important part of VQ-HPS, is pre-trained on AMASS [29] is a great advantage of our approach. Indeed, we can leverage large amounts of body mesh data not paired with images for training. This is particularly interesting because many body motion data can be generated using animation software or generative models. We could finetune the Mesh-VQ-VAE depending on the target application, with uncommon body shapes [82] or extreme poses [83, 84].

FastMETRO slightly takes the lead in global metrics on the EMDB benchmark when training on COCO. The advantage VQ-HPS had when training with scarce data is less important here. Indeed, COCO has as many images as EMDB or 3DPW, but the diversity is much higher than for video datasets, where there exist important correlations between different images [33].

5.4 Training on large-scale datasets

Following the standard practice [1, 2, 3], we train VQ-HPS on Human3.6M [77], MPI-INF-3DHP [86], COCO [79], and MPII [87]. For this experiment, we use the same version of VQ-HPS as in Sec. 5.3. We evaluate VQ-HPS on 3DPW [30] and EMDB [31] without finetuning on the 3DPW training set. We only compare VQ-HPS with SOTA models using the same backbone and the same datasets for a fair comparison. We take results from the papers or use the provided implementations and checkpoints for other methods. Note that the results on [3, 85] differ from the papers because we do not finetune the models on the 3DPW [30] dataset before testing. Recent methods using a different backbone such as a Vision Transformer (ViT) [49], as well as additional datasets [88, 60, 61, 89] may obtain better results. However, the comparison would not be fair. Results are shown in Tab. 2.

Table 2: **Comparison to the SOTA methods.** We evaluate VQ-HPS trained on a mix of datasets without finetuning on 3DPW (see Sec. 5.1) with standard metrics on 3DPW and EMDB and compare them to the state of the art. On the left part, methods use a ResNet-50 backbone. On the right, models use an HRNet backbone, except TokenHMR [76] that uses a ViT [49] backbone and additional data. All results are given in mm.

(a) ResNet-50 backbone

Method	PVE ↓	MPJPE ↓	PA-MPJPE ↓
GraphCMR [19]	-	-	70.2
I2LMeshNet [63]	-	93.2	58.6
FastMETRO-S [3]	129.4	112.6	68.9
HMR [1]	-	130.0	81.3
SPIN [39]	116.4	96.9	59.2
PyMAF [42]	110.1	92.8	58.9
ROMP [47]	105.6	89.3	53.5
DSR [48]	105.8	91.7	54.1
PARE [41]	99.7	82.9	52.3
VQ-HPS (ours)	**93.6**	**79.1**	**50.4**

(b) HRNet backbone

Method	3DPW			EMDB		
	PVE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓	MPJPE ↓	PA-MPJPE ↓
FastMETRO-L [3]	121.6	109.0	65.7	119.2	108.1	72.7
ROMP [47]	103.1	85.5	54.9	134.9	112.7	75.2
PARE [41]	97.9	82.0	50.9	133.2	113.9	72.2
Virtual Marker [85]	93.8	80.5	48.9	-	-	-
CLIFF [2]	87.6	73.9	46.4	122.9	103.1	68.8
TokenHMR [76]	*88.1*	*76.2*	*49.3*	*124.4*	*102.4*	*67.5*
VQ-HPS (ours)	**84.8**	**71.1**	**45.2**	**112.9**	**99.9**	**65.2**

Table 3: **Ablation study.** We perform several ablations on the VQ-HPS architecture and training process on the 3DPW dataset. All results are given in mm.

Method	PVE ↓	MPJPE ↓	PA-MPJPE ↓
VQ-HPS	176.6	152.0	91.8
SMPL	199.8	171.8	99.3
3D loss	220.3	194.9	144.1
No reprojection	183.9	158.4	95.6

VQ-HPS outperforms all other methods on all 3 metrics, being parametric [1, 39, 42, 47, 48, 41, 2] or non-parametric [3, 85, 19, 63]. Note that there is a large gap between the performance of FastMETRO and Virtual Marker in Tab. 2 and the reported results in the original papers. This is because we do not perform finetuning on 3DPW. The authors of FastMETRO acknowledge that methods regressing 3D vertices such as [3, 24, 23, 85] perform poorly on data significantly different from the training set[1]. This limitation of non-parametric methods was also described in [24].

Even though TokenHMR's backbone is more powerful and the method is trained on additional 2D datasets, VQ-HPS still outperforms TokenHMR [76] on the 3DPW and EMDB datasets. When additionally incorporating Bedlam [60] in the training set, TokenHMR takes the lead on the 3DPW dataset as it obtains 84.6, 71.0, and 44.3 mm for the PVE, MPJPE, and PA-MPJPE, respectively, but VQ-HPS remains competitive using much less training data and a less powerful backbone. On the EMDB dataset, TokenHMR using Bedlam for training obtains 109.4, 91.7, and 55.6 mm for the PVE, MPJPE, and PA-MPJPE, respectively.

5.5 Ablation study

We ablate VQ-HPS architecture and training scheme and present the results in Tab. 3. We train and test on the 3DPW [30] dataset for these experiments. Note that for faster experiments, we use early stopping with patience of 10 epochs, which makes the training stop much earlier. The first line of Tab. 3 shows the updated VQ-HPS results.

"SMPL" means predicting the SMPL parameters instead of using our proposed discrete latent representation. For obtaining the final 3D prediction, the SMPL model is used instead of the Mesh-VQ-VAE decoder. The performance gap shows that using similar architecture, predicting the discrete latent representation instead of the SMPL parameters yields improved performance.

The ablation "3D loss" replaces the cross-entropy loss with the PVE $\mathcal{L}_{3D} = ||V - \hat{V}||_2$ where V is the ground truth mesh vertices and $\hat{V}$ the final prediction. Given the huge decrease in performance (see Tab. 3), we conclude that cross-entropy is a good alternative to 3D losses used in all prior works such as [1, 2, 3]. Training VQ-HPS with the PVE produces sequences of indices corresponding to non-anthropomorphic results, which recall results obtained with FastMETRO in Fig. 4 when training on limited data.

[1] https://github.com/postech-ami/FastMETRO/issues/13

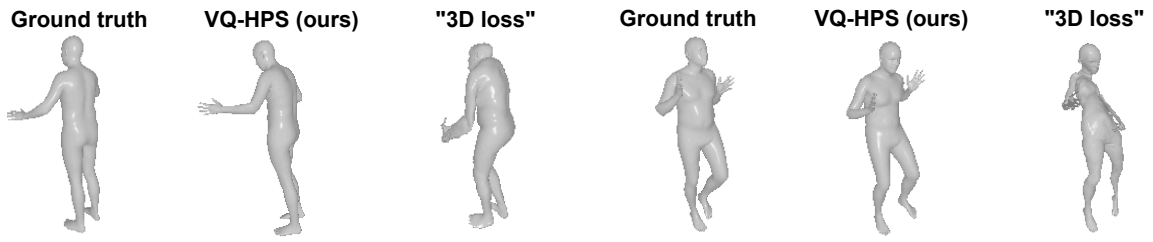

Figure 5: **Ablation study.** Effect of the "3D loss" ablation. We can see that replacing the cross-entropy with a PVE loss produces unnatural poses, showing interest in the classification-based approach.

"No reprojection" means that we do not compute the reprojection error. This mostly increases the error in PVE and MPJPE, which was expected since the PA-MPJPE is only related to the canonical mesh, and the reprojection loss is not used to train the canonical mesh predictor. However, it still has an impact since the canonical mesh prediction is conditioned on the predicted rotation.

6 Conclusion

In this work, we proposed Mesh-VQ-VAE, an autoencoder architecture providing a discrete latent representation of registered human meshes. This discrete representation allowed us to tackle the HPSE problem from a classification perspective, avoiding the limitations of parametric and non-parametric HPSE methods described in Sec. 1. We also introduced VQ-HPS, a Transformer-based model for solving the proposed HPSE classification problem.

While trained using the cross-entropy loss, VQ-HPS significantly outperforms state-of-the-art methods when trained on scarce data and shows promising performance on large-scale datasets. The classification-based approach exploits the discriminative power of Transformers. It addresses several known problems of non-parametric approaches, such as the plausibility of results and the lack of generalization when training on large-scale datasets without finetuning on the target domain. Comparisons to [3] as well as our ablation study (Sec. 5.5) showed the superiority of the classification-based approach compared to parametric and existing non-parametric approaches when using similar architectures.

Failure cases are discussed in Appendix E. Future works may include the use of a SOTA backbone [49, 90] or additional datasets [89, 60] for achieving better performance. Extensions of the classification-based approach may also be explored for other types of registered meshes, such as human hands or faces. We also believe that multimodal applications involving text and 3D humans [91, 92] would benefit from the Mesh-VQ-VAE representation as it can be considered a language.

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. This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

References

[1] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[2] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606. Springer, 2022.

[3] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, pages 342–359. Springer, 2022.

[4] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 41(4), 2022.

[5] Andrew Feng, Samuel Shin, and Youngwoo Yoon. A tool for extracting 3d avatar-ready gesture animations from monocular videos. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (ACM MIG)*, pages 1–7, 2022.

[6] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.

[7] Moritz Einfalt, Dan Zecha, and Rainer Lienhart. 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)*, pages 446–455. IEEE, 2018.

[8] Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *ACM International Conference on Multimedia (ACM MM)*, pages 374–382, 2019.

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015.

[10] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3), 2005.

[11] Ahmed A. A. Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human representation. In *European Conference on Computer Vision (ECCV)*, pages 568–585. Springer, 2022.

[12] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[13] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6184–6193, 2020.

[14] 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 *European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016.

[15] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017.

[16] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499, 2021.

[17] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14783–14794, 2023.

[18] Enric Corona, Gerard Pons-Moll, Guillem Alenyà, and Francesc Moreno-Noguer. Learned vertex descent: a new direction for 3d human model fitting. In *European Conference on Computer Vision (ECCV)*, pages 146–165. Springer, 2022.

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019.

[20] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, pages 769–787. Springer, 2020.

[21] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3d human shape and pose from dense body parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(5):2610–2627, 2020.

[22] Dongkai Wang and Shiliang Zhang. 3d human mesh recovery with sequentially global rotation estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14953–14962, 2023.

[23] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021.

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30:5998–6008, 2017.

[26] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15143–15155, 2023.

[27] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems (NIPS)*, 30:6306–6315, 2017.

[28] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *Advances in neural information processing systems (NIPS)*, 33:9251–9262, 2020.

[29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019.

[30] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 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, 2018.

[31] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14632–14643, 2023.

[32] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3d human pose and shape estimation: A sparse constrained formulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11457–11466, 2021.

[33] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021.

[34] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018.

[35] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894, 2019.

[36] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, pages 572–589. Springer, 2022.

[37] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14725–14737, 2023.

[38] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Asian Conference on Computer Vision (ACCV)*, 2020.

[39] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019.

[40] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263, 2020.

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

[42] 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021.

[43] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pages 284–300. Springer, 2020.

[44] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1964–1973, 2021.

[45] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019.

[46] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(9):4490–4504, 2021.

[47] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021.

[48] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J Black. Learning to regress bodies from images using differentiable semantic rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11250–11259, 2021.

[49] 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. In *International Conference on Learning Representations (ICLR)*, 2021.

[50] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168, 2023.

[51] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems (NIPS)*, 36, 2024.

[52] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1611–1620, 2023.

[53] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. 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)*, pages 16094–16104, 2021.

[54] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11605–11614, 2021.

[55] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems (NIPS)*, 33:20496–20507, 2020.

[56] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8781–8791, 2023.

[57] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. 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)*, pages 4779–4789, 2023.

[58] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017.

[59] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021.

[60] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 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, 2023.

[61] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021.

[62] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2299–2307, 2022.

[63] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 752–768. Springer, 2020.

[64] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022.

[65] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(5):1146–1161, 2019.

[66] Grégory Rogez and Cordelia Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. *Advances in neural information processing systems (NIPS)*, 29:3108–3116, 2016.

[67] Isaac Cohen and Hongxia Li. Inference of human postures by classification of 3d human body shape. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 74–81, 2003.

[68] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *European Conference on Computer Vision (ECCV)*, pages 852–863. Springer, 2012.

[69] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3056–3065, 2017.

[70] Pornthep Sarakon, Theekapun Charoenpong, and Supiya Charoensiriwath. Face shape classification from 3d human data by using svm. In *Biomedical Engineering International Conference*, pages 1–5. IEEE, 2014.

[71] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Head pose estimation: Classification or regression? In *International Conference on Pattern Recognition (ICPR)*, pages 1–4. IEEE, 2008.

[72] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision (ECCV)*, pages 417–435. Springer, 2022.

[73] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11050–11059, 2022.

[74] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023.

[75] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2321–2330, 2023.

[76] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. *arXiv preprint arXiv:2404.16752*, 2024.

[77] 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 (PAMI)*, 36(7):1325–1339, 2013.

[78] Tianshu Zhang, Buzhen Huang, and Yangang Wang. 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, 2020.

[79] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[80] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[81] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015.

[82] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020.

[83] Guénolé Fiche, Vincent Sevestre, Camila Gonzalez-Barral, and Simon Leglaive. Swimxyz: A large-scale dataset of synthetic swimming motions and videos. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG)*, 2023.

[84] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13053–13064, 2022.

[85] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–543, 2023.

[86] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[87] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[88] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3925–3935, 2023.

[89] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*, pages 557–577. Springer, 2022.

[90] Matthieu Armando, Salma Galaaoui, Fabien Baradel, Thomas Lucas, Vincent Leroy, Romain Brégier, Philippe Weinzaepfel, and Grégory Rogez. Cross-view and cross-pose completion for 3d human understanding. *arXiv preprint arXiv:2311.09104*, 2023.

[91] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision (ECCV)*, pages 346–362. Springer, 2022.

[92] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Posegpt: Chatting about 3d human pose. *arXiv preprint arXiv:2311.18836*, 2023.

Supplementary material

A Implementation details

In this section, we provide the details of the implementation of Mesh-VQ-VAE and VQ-HPS. More information on these models can be found in the main paper (see Sec. 4.1 and Fig. 2).

Mesh-VQ-VAE. The Mesh-VQ-VAE architecture (see Fig. 2) is adapted from the fully convolutional mesh autoencoder of [1]. This model encodes a mesh to a latent representation $z \in \mathbb{R}^{N \times L}$ with $N = 54$ and $L = 9$. The quantization step is performed with a dictionary of size $S = 512$.

Feature extractors. Our CNN backbones are ResNet-50 [2] pre-trained on ImageNet [3] to provide a fair comparison with previous methods. For the canonical mesh prediction, we remove the last fully connected layer to obtain features $X_{mesh} \in \mathbb{R}^{H \times W \times C}$ with $H = W = 7$, $C = 2048$. For the rotation and camera prediction, we keep the full Resnet-50 to obtain features $X_{rot} \in \mathbb{R}^{C}$.

Rotation and camera predictors. Inspired by [4], the network for predicting the rotation and the mesh consists of a multilayer perceptron (MLP) regression module composed of two fully connected layers with 1024 neurons following the CNN backbone (see Fig. 2). The image feature $X_{rot} \in \mathbb{R}^{C}$ is concatenated with the flattened pose $p \in \mathbb{R}^{17*3}$ before being fed to the MLP. The rotation is predicted in the 6d-rotation format [5], and p is initialized with the SMPL T-pose.

Latent canonical mesh regressor. The Transformer follows the original Transformer architecture [6]. Its hidden dimension is $D = 512$, and all MLPs in the encoder and decoder have a hidden size of 1024. It has 5.3 million parameters (7.6 million for the version used in the COCO [7] and large-scale training). We use sinusoidal positional encoding for the input tokens.

B Additional qualitative results

Additional comparisons with other methods trained on little data are provided in Fig. 1 to complement the results of Fig. 4.

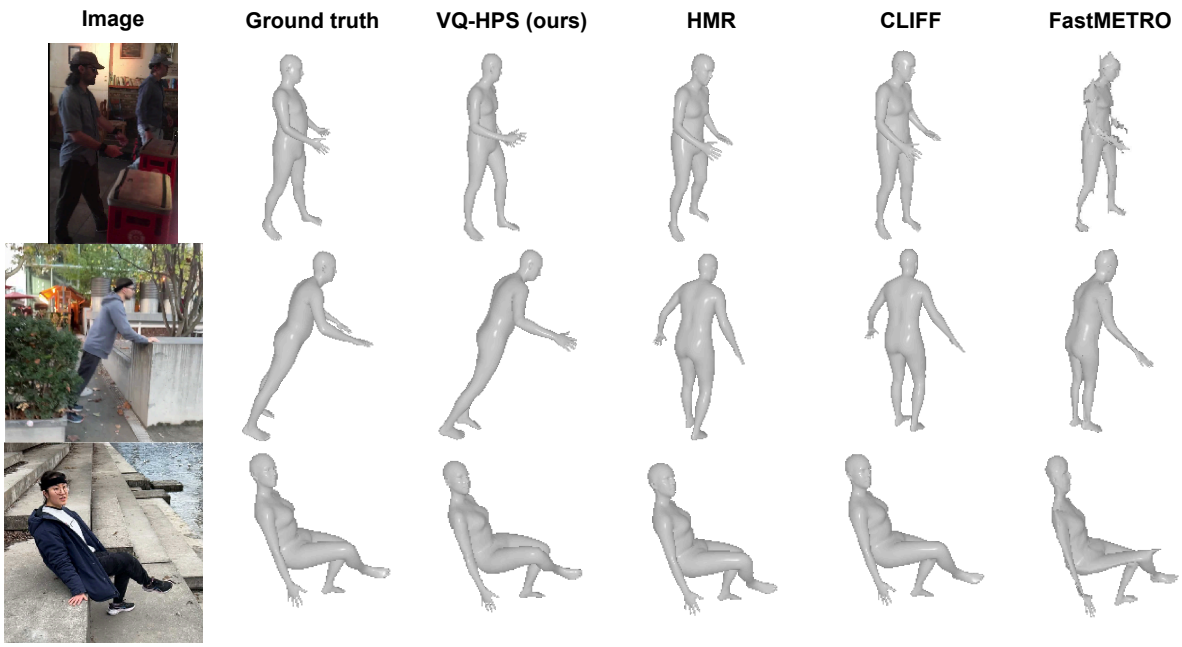

Figure 1: **Additional comparisons.** We compare our method with HMR, CLIFF, and FastMETRO-S on 3DPW trained on 3DPW (first row) and EMDB trained on EMDB.

Figure 2: **Qualitative results.** We visualize results obtained with VQ-HPS on the 3DPW dataset.

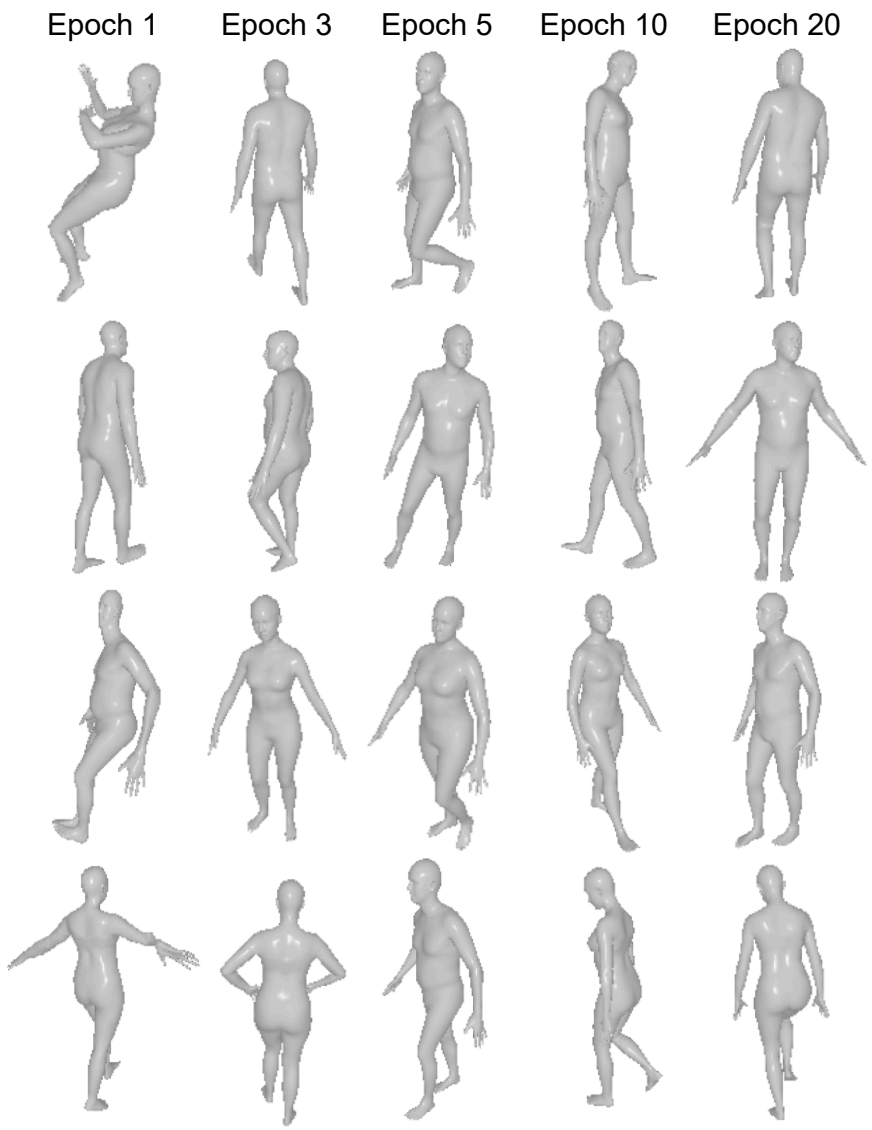

Figure 3: Visualization of random validation samples during the training process on EMDB. Meshes are smooth from the first epoch and become anthropomorphic in about 5 epochs.

We also provide some qualitative results obtained with VQ-HPS using an HRNet backbone as evaluated in Tab. 2) are shown in Fig. 2.

C Visualization of predictions during training

We visualize validation samples when training on the EMDB dataset (see Sec. 5.3) in Fig. 3. The meshes are smooth from the first epoch and become anthropomorphic in about 5 epochs. As discussed in the paper, we believe this is due to multiple factors. First, the Mesh-VQ-VAE, whose decoder is essential to VQ-HPS, was pre-trained on large-scale human motion datasets and is frozen afterward. This pre-training is probably a regularization that reduces the labeled data needed for learning to solve the HPSE task. Second, VQ-HPS learns a distribution over discrete indices for producing anthropomorphic meshes. Learning a distribution over 54 discrete indices is easier than learning the structure from 6890 3D coordinates.

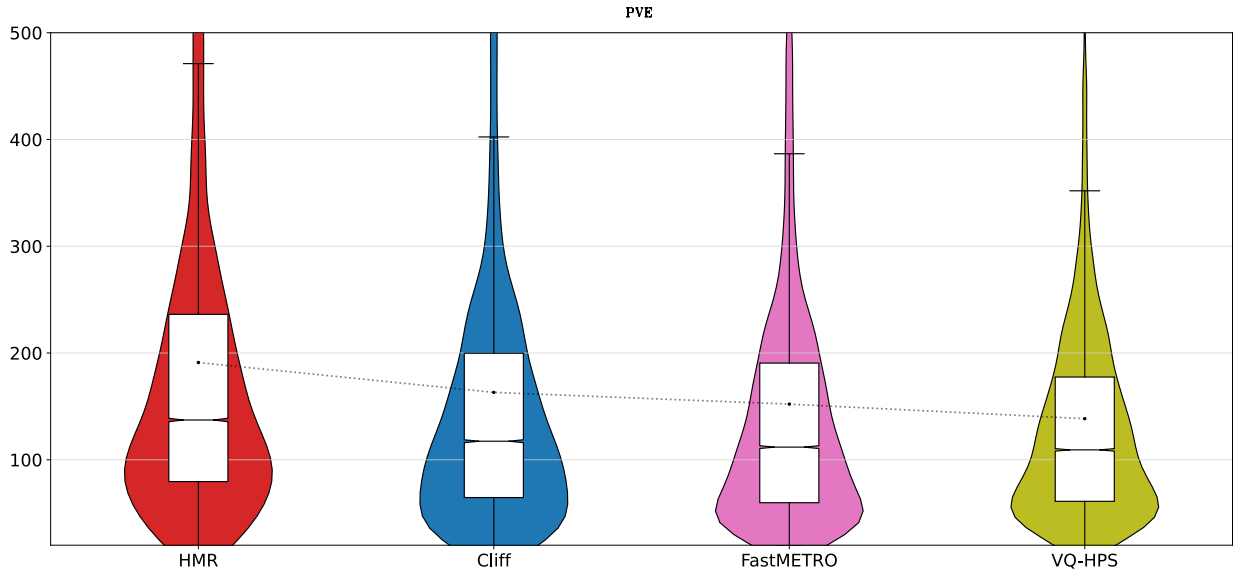

Figure 4: **Distribution of the PVE.** We study the distribution of the per-vertex error for all 4 methods on the testing set of EMDB. Black dots represent the mean. The box plots give the 1st and 3rd quartiles, as well as the median value, with a notch for the confidence interval around the median value. The whiskers extend from the box to the farthest data point, lying within 1.5x the interquartile range (IQR) from the box. We also draw the violin plot to visualize the distribution of the error.

D Analysis of the error

We analyze the distribution of the per-vertex error (PVE, see Sec. 5.2) on EMDB1 when models are trained on EMDB2 (see Sec. 5.3, and quantitative results in Tab. 1) in Fig. 4. We can see that VQ-HPS outperforms other methods in many ways. First, its mean, median, 1st, and 3rd quartiles are lower than other methods. The distance between the 1st and 3rd quartiles is much smaller for VQ-HPS than for other methods. Non-parametric methods have a very high concentration of samples with low error, while parametric methods have a more uniform distribution. Another significant advantage of VQ-HPS is its many fewer outliers with high errors.

E Failure cases

Failure cases are shown in Fig. 5. When training on scarce data (experiments of Sec. 5.3), low visibility and unusual poses lead to outliers with completely different poses and body orientations. There is a clear improvement when training on large-scale datasets (see Sec. 5.4), but there are still some failure cases. In the first image, the model estimates the pose of the wrong person. In the second image, a very unusual pose leads to a non-anthropomorphic prediction (especially for the left arm). Finally, VQ-HPS sometimes makes global orientation mistakes for unusual poses. Potential improvements for avoiding such failure cases would be improving the feature extractors [8] or using additional data with more unusual poses [9], as recent works [10] highlighted the pivotal role of backbones and data for obtaining better performance.

F Estimating body shapes

We propose to evaluate the potential of VQ-HPS for estimating body shapes. As demonstrated in prior works, the most important for estimating accurate body shapes is the training data [9]. We finetune VQ-HPS (see Sec. 5.4, "ResNet-50 backbone" in Tab. 2) on 5% of Bedlam [9], a synthetic dataset with diverse body shapes. In Fig. 6, we compare the results obtained on SSP-3D [11], a dataset of real images with challenging body shapes, before and after finetuning VQ-HPS on Bedlam. The clear improvement suggests that VQ-HPS could be used for body shape estimation if trained on appropriate data.

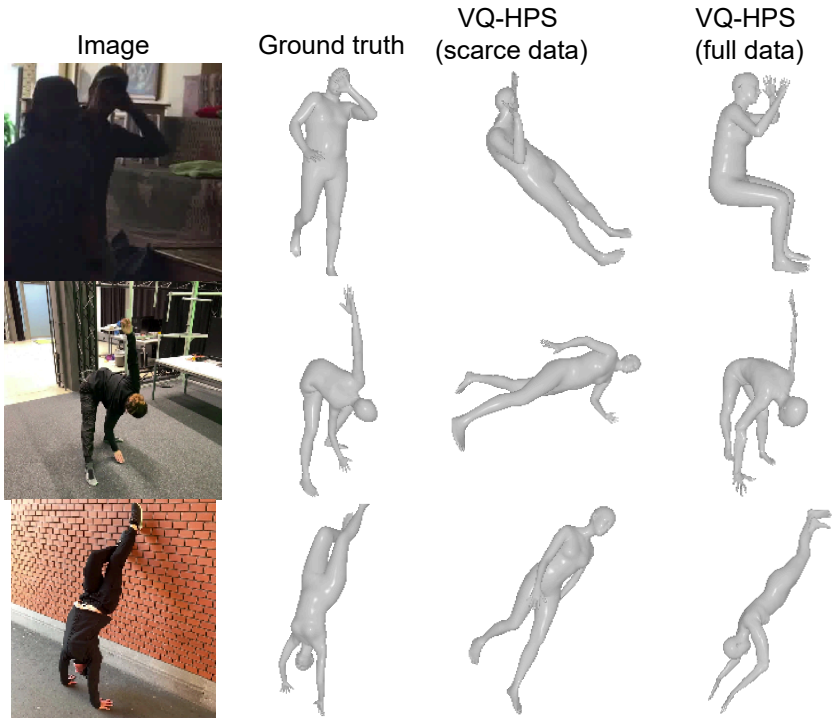

Figure 5: **Failure cases.** We study the failure cases when training on scarce data and large-scale datasets.

Figure 6: **Estimating body shapes.** We evaluate VQ-HPS qualitatively on SSP-3D before and after finetuning on a dataset with diverse body shapes.

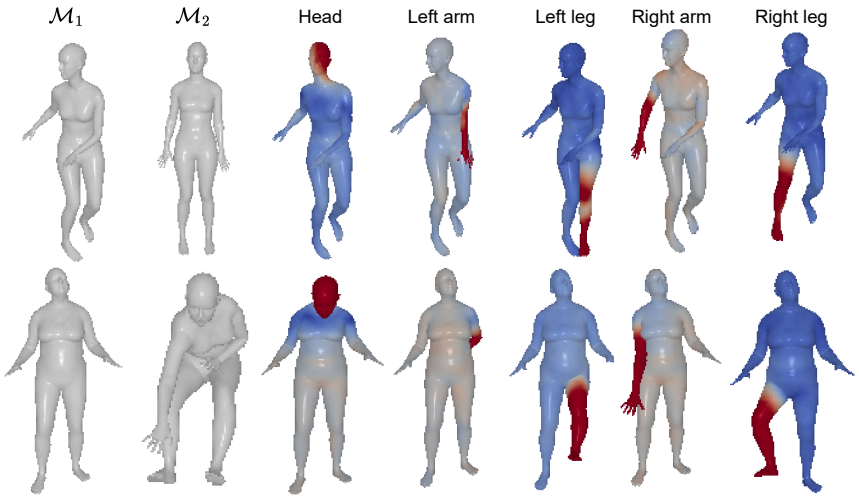

Figure 7: **Exchanging body parts.** We exchange body parts between the two meshes in the latent space after manually identifying the indices responsible for each body part. The color map shows the distance between $\mathcal{M}_1$ and the reconstruction.

Figure 8: **Exchanging the torso.** We exchange the torsos between the two meshes in the latent space.

G Visualization of the Mesh-VQ-VAE

The main paper (see Sec. 4.1) explains that the Mesh-VQ-VAE is fully convolutional. Hence, the latent space preserves the spatial structure of the mesh. We manually identify the body part associated with each index by visualizing the reconstruction after randomly modifying each. We propose to visualize this property by exchanging body parts between

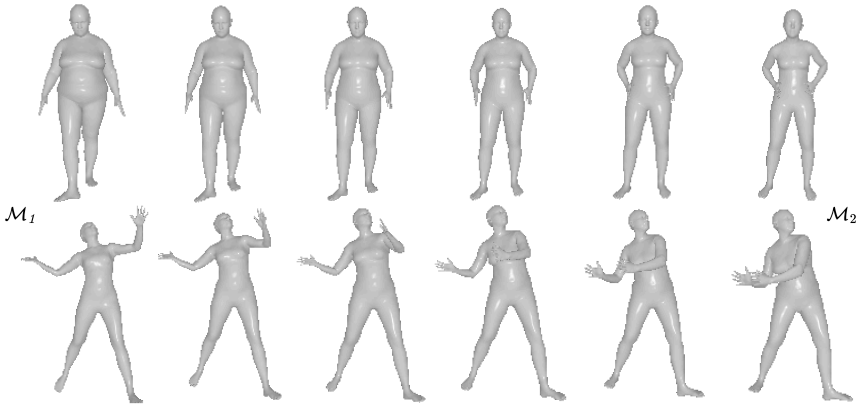

Figure 9: **Interpolation in the latent space of Mesh-VQ-VAE.** Note that the latent space encapsulates both pose and shape.

different meshes. Specifically, given two meshes $\mathcal{M}_1$ and $\mathcal{M}_2$, we encode both meshes and decode $\mathcal{M}_1$ after replacing the indices of a given body part by $\mathcal{M}_2$. Results are shown in Fig. 7.

We can also modify the torso. To visualize it easier, we provide qualitative results for individuals with different body shapes in Fig. 8. Note that this slightly differs from modifying the body shape, as the arms and legs are unchanged.

Finally, in Fig. 9, we show that we can interpolate between meshes $\mathcal{M}_1$ and $\mathcal{M}_2$ in the latent space. Specifically, we do a linear interpolation between the corresponding continuous latent representations z_1 and z_2, which are then quantized and decoded to obtain intermediate meshes.

Supplementary References

[1] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *Advances in neural information processing systems (NIPS)*, 33:9251–9262, 2020.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015.

[4] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[5] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30:5998–6008, 2017.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[8] Matthieu Armando, Salma Galaaoui, Fabien Baradel, Thomas Lucas, Vincent Leroy, Romain Brégier, Philippe Weinzaepfel, and Grégory Rogez. Cross-view and cross-pose completion for 3d human understanding. *arXiv preprint arXiv:2311.09104*, 2023.

[9] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 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, 2023.

[10] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. *Advances in Neural Information Processing Systems (NIPS)*, 35:26034–26051, 2022.

[11] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020.