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] Francesc Moreno-Noguer[3]

[1]CentraleSupélec, IETR UMR CNRS 6164, France [2]Inria, UGA, CNRS, LJK, France [3]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

The 18th European Conference on Computer Vision (ECCV), Milano, Italy, 2024

1. Introduction

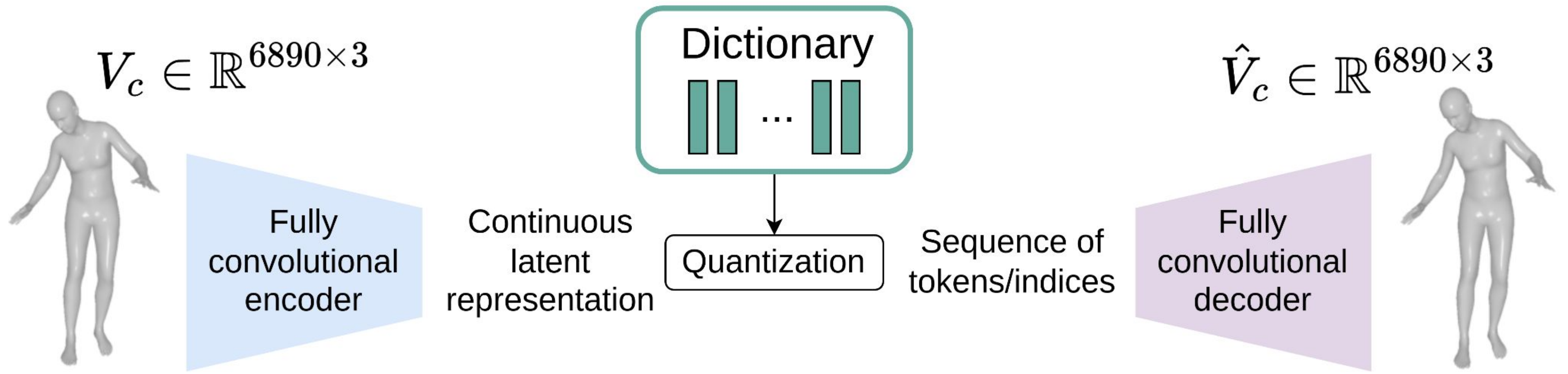
Human mesh recovery (HMR) aims to regress a 3D human body model from an image.

We propose to address the HMR problem as a **classification task**:

- We introduce the **Mesh-VQ-VAE,** an autoencoder that learns a discrete token representation of human meshes.
- We propose **VQ-HPS**, a Transformer-based model to solve HMR in the quantized latent space of the Mesh-VQ-VAE.

2. Mesh-VQ-VAE

Mesh-VQ-VAE is a **fully convolutional** mesh autoencoder [1] with a **quantized latent space** akin to VQ-VAE [2].

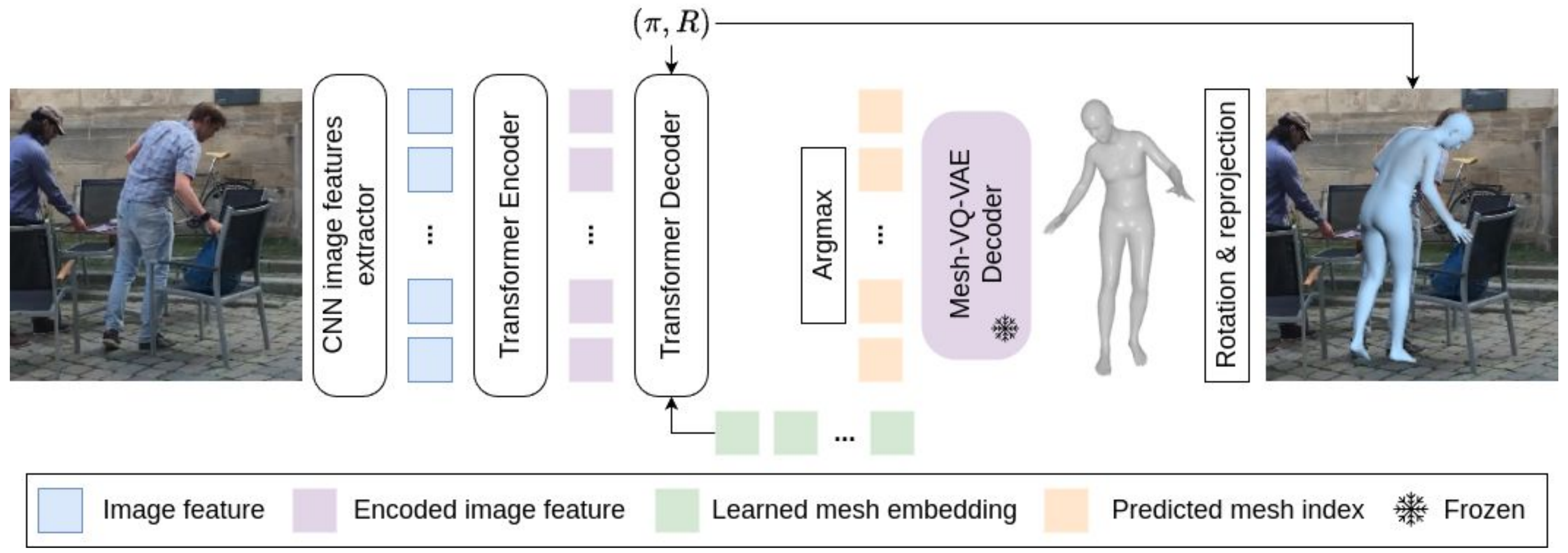


- A human mesh is represented with **54 human mesh tokens**.
- We minimize the reconstruction error on the AMASS dataset.
- Tokens are **easy-to-process** for Transformers.
- Quantization acts as a **prior**.

[1] Zhou, Yi, et al. "Fully convolutional mesh autoencoder using efficient spatially varying kernels." NeurIPS. 2020
[2] Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." NeurIPS. 2017

3. VQ-HPS

VQ-HPS is a Transformer encoder-decoder that predicts human meshes as a **sequence of discrete tokens**.



- Once predicted, the sequence of tokens is decoded with the **pre-trained decoder** of the Mesh-VQ-VAE.
- We train VQ-HPS on a **mixture of datasets** annotated with pseudo-ground truth meshes.
- One single training loss function: the **cross-entropy**.

4. Quantitative results

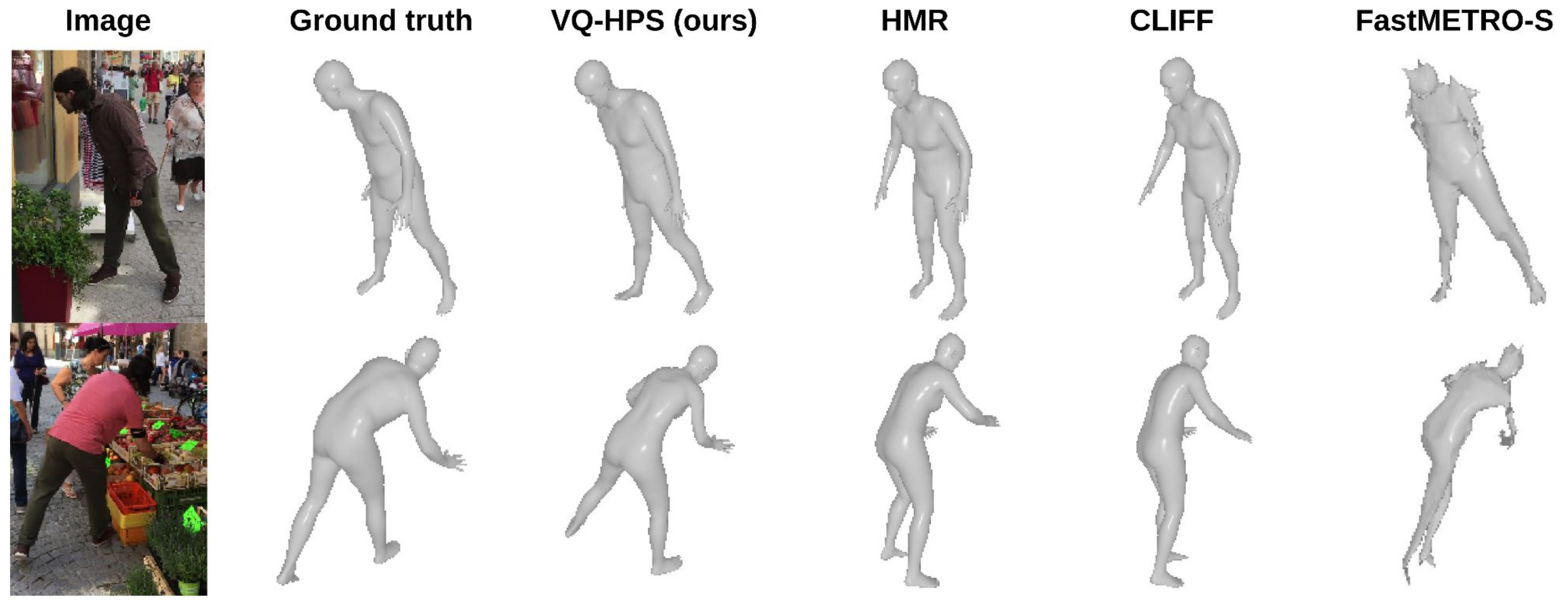
We evaluate VQ-HPS on in-the-wild datasets **without fine-tuning** on the 3DPW training set.

Method	3DPW PVE ↓	3DPW MPJPE ↓	3DPW PA-MPJPE ↓	EMDB PVE ↓	EMDB MPJPE ↓	EMDB PA-MPJPE ↓
FastMETRO-L	121.6	109.0	65.7	<u>119.2</u>	108.1	72.7
ROMP	103.1	85.5	54.9	134.9	112.7	75.2
PARE	97.9	82.0	50.9	133.2	113.9	72.2
Virtual Marker	93.8	80.5	48.9	-	-	-
CLIFF	<u>87.6</u>	<u>73.9</u>	<u>46.4</u>	122.9	<u>103.1</u>	<u>68.8</u>
TokenHMR	*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**

VQ-HPS obtains state-of-the-art performance in HMR.

5. Training with scarce data

We compare VQ-HPS to other methods when training only on the 3DPW training set.



Quantization allows to obtain **accurate and realistic results even with little training data**.

6. Perspectives

Future work may include:

- Developing quantized representation and similar approaches for **hands, faces, or full-body reconstruction**.
- Quantized representation can be linked to many **different modalities**, such as text or audio.

Following this work we proposed **MEGA**:

- Uses the quantized human mesh representation of Mesh-VQ-VAE.
- Single and multi-output HMR with **masked generative modeling**.
- SOTA results in **single and multi-output** HMR.
- More info: https://g-fiche.github.io/research-pages/mega/