MEGA: Masked Generative Autoencoder for Human Mesh Recovery

Guénolé Fiche[1,2] Simon Leglaive[1] Xavier Alameda-Pineda[3] Francesc Moreno-Noguer[4,5]

[1]CentraleSupélec [2]Naver Labs Europe [3]Inria Grenoble [4]Institut de Robòtica i Informàtica Industrial [5]Amazon

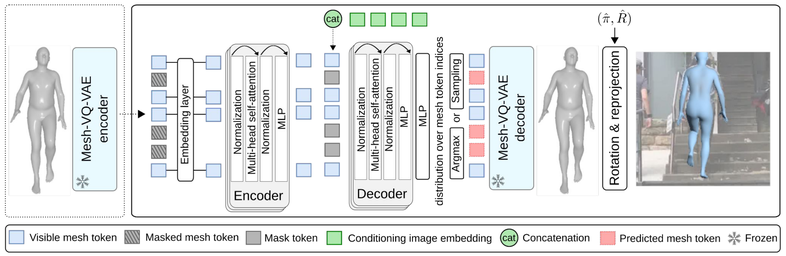
1. Motivation

Human mesh recovery from a single image is highly ambiguous, as an infinite set of predictions match the 2D observations equally.

- Single-output methods overlook the ambiguity and focus on high accuracy and efficiency.
- Multi-output methods propose diverse solutions given an image but are less accurate when making a single prediction.

2. MEGA

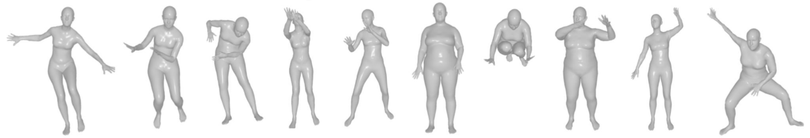
Masked generative modeling enables efficient single- and multi-output predictions by leveraging a tokenized latent representation of human meshes.



MEGA relies on Mesh-VQ-VAE, an autoencoder that represents a human mesh as a sequence of 54 tokens.

3. Training procedure

- MEGA is first pre-trained for masked token prediction on a large-scale motion capture dataset without paired images → random human mesh generation.



- Then, MEGA is fine-tuned for the same task but with additional conditioning from RGB images → HMR.
- In both training stages, only the cross-entropy loss is used.

4. Deterministic mode

MEGA can be used as a single-output HMR model by predicting all tokens simultaneously given an image and a fully-masked sequence of human mesh tokens.

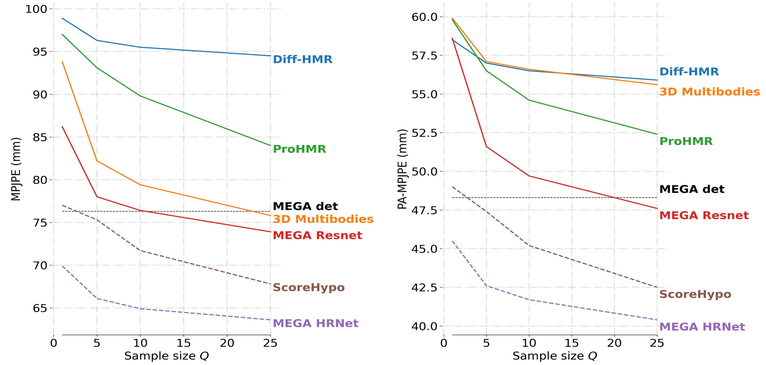
Method	Backbone	3DPW PVE ↓	3DPW MPJPE ↓	3DPW PA-MPJPE ↓	EMDB PVE ↓	EMDB MPJPE ↓	EMDB PA-MPJPE ↓
FastMETRO [13]	HRNet-w64	121.6	109.0	65.7	119.2	108.1	72.7
PARE [42]	HRNet-w32	97.9	82.0	50.9	133.2	113.9	72.2
Virtual Marker [61]	HRNet-w48	93.8	80.5	48.9	-	-	-
CLIFF [52]	HRNet-w48	87.6	73.9	46.4	122.9	103.1	68.8
VQ-HPS [24]	HRNet-w48	84.8	71.1	45.2	112.9	99.9	65.2
MEGA (ours)	HRNet-w48	**81.6**	**68.5**	**44.1**	**107.9**	**90.5**	**58.7**
linear masking	HRNet-w48	86.5	72.6	45.9	118.7	100.1	63.3
full mask	HRNet-w48	81.8	**68.5**	44.4	110.3	92.7	59.2
w/o pre-training + full mask	HRNet-w48	84.1	70.5	46.2	113.9	95.9	62.0
HMR2.0 [27]	ViT-H	84.1	70.0	44.5	120.1	97.8	61.5
TokenHMR† [22]	ViT-H	88.1	76.2	49.3	124.4	102.4	67.5
TokenHMR∓ [22]	ViT-H	84.6	71.0	44.3	109.4	**91.7**	55.6
MEGA (ours)	ViT-H	**80.0**	**67.5**	**41.0**	**108.6**	92.4	**52.5**

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 [7]. Methods in italic below the row "MEGA" indicate the results of the ablation study.

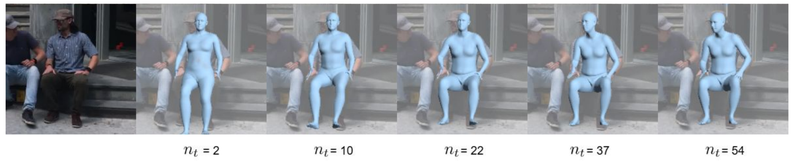
We compare MEGA to single-output methods on in-the-wild datasets without finetuning on the 3DPW dataset.
MEGA is also SOTA on the occlusion dataset 3DPW-OCC despite not being designed specifically to handle occlusions.

5. Stochastic mode

We can generate diverse human meshes iteratively. MEGA is then a multi-output HMR method.



The iterative process can be visualized as follows:



The diversity in prediction can be interpreted as a measure of the per-vertex uncertainty: