TokenHMR:
Advancing Human Mesh Recovery with a Tokenized Pose Representation

Sai Kumar Dwivedi[1,*], Yu Sun[2,*], Priyanka Patel[1], Yao Feng[1,2,3], Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, [2] Meshcapade, [3] ETH Zurich

(*Equal Contribution)

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS · meshcapade · ETH zürich

CVPR · SEATTLE, WA JUNE 17-21, 2024

Goal

Observations
1. Most 3D human pose estimation methods train on real images with 2D keypoints and/or 3D pseudo ground-truth.
2. This helps with robustness and generalization.
3. Methods trained on such data exhibit good image alignment but poor 3D accuracy.

Proposal

TokenHMR → Generalizes in-the-wild · Without drop in 3D accuracy

Problem

a) Optimization process for generating pseudo ground-truth

b) Projection of 3D joints to 2D using simplified camera model

Analysis-I

HRM2.0 Predicted Bodies + HMR2.0 Predicted Camera

BEDLAM Ground-truth Bodies + HMR2.0 Predicted Camera

Front-view rendering · Side-view rendering · Projected 2D keypoints

	Validation Seq. of BEDLAM	
	PCK-0.5	PCK-1.0
BEDLAM[1] Ground-truth Bodies	0.66	0.86
HMR2.0[2] Predicted Bodies	**0.78**	**0.88**

[1] Black et al. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. CVPR, 2023
[2] Goel et al. Humans in 4D: Reconstructing and Tracking Humans with Transformers. ICCV, 2023

Analysis-II

BEDLAM-GT bodies in HMR2.0 camera

Front-view rendering · Side-view rendering · Projected 2D keypoints

Keeping 2D alignment, how much can 3D vary?

Similar 2D Alignment, Different 3D Pose

Front-view rendering · Side-view rendering · Projected 2D keypoints

TALS Loss

Previous Methods

Trade-off between generalization and bias

2D Datasets & 3D Pseudo-GT → Do Not Use → Poor Generalization

2D Datasets & 3D Pseudo-GT → Use → Pose/Camera Bias

TALS — Threshold Adaptive Loss Scaling

Pseudo-GT

$$\mathcal{L}_{\theta_{pGT}} = \begin{cases} \|\theta - \theta_g\|^2 & \text{if } \mathcal{L}_{\theta_{pGT}} > \varepsilon_\theta \\ \alpha_\theta \cdot \|\theta - \theta_g\|^2 & \text{otherwise} \end{cases}$$

2D Keypoints

$$\mathcal{L}_{J_{2D_{pGT}}} = \begin{cases} |J_{2D} - J_{2D_g}| & \text{if } \mathcal{L}_{J_{2D_{pGT}}} > \varepsilon_{J_{2D}} \\ \alpha_{J_{2D}} \cdot |J_{2D} - J_{2D_g}| & \text{otherwise} \end{cases}$$

$\mathcal{L}_{\theta, J_{2D}} > \varepsilon$ → Helps in Generalization

$\mathcal{L}_{\theta, J_{2D}} < \varepsilon$ → Prevents Bias

Need for Tokenization

Problem: With TALS loose fitting, predicting 3D from 2D becomes more ambiguous, hence we need an unbiased prior of valid poses.

Gaussian Mixture Model · Variational AutoEncoder (VAE)

Biased towards frequently occurring poses

Encoder → Codebook → Decoder

Pose (θ) → 1 9 9 4 ... 6 → Pose (θ)

Discrete SMPL Pose Tokens

Trained with AMASS, ICCV'19 MOYO, CVPR'23

Codebook of Valid Poses

Summary

Overall Architecture

Backbone → Transformer blocks → Global Orientation / Pose Classification / Shape / Camera → Softmax × Codebook → Decoder → Pose (Frozen)

Key Ideas
- TALS loss: Avoid overfitting to 2D keypoints and 3D pseudo-GT
- Tokenization: Unbiased prior of valid poses.

Results on 3D Datasets

Training Datasets	Method	EMDB (ICCV 2023)			3DPW (ECCV 2018)		
		MVE	MPJPE	PA-MPJPE	MVE	MPJPE	PA-MPJPE
SD	HybrIK (CVPR 2023)	122.2	103.0	65.6	94.5	80.0	48.8
SD	CLIFF (ECCV 2022)	122.9	103.1	68.8	81.2	69.0	43.0
SD	HMR2.0 (ICCV 2023)	120.1	97.8	61.5	84.1	70.0	44.5
BL	TokenHMR	104.2	88.1	49.8	86.0	70.5	43.8
SD + ITW + BL	HMR2.0	120.7	99.3	62.8	88.4	77.4	47.4
SD + ITW + BL	HMR2.0 + TALS	115.7	96.7	58.5	89.6	73.5	46.8
SD + ITW + BL	HMR2.0 + Token	116.1	95.6	62.2	86.6	75.0	48.0
SD + ITW + BL	HMR2.0 + TALS + VPoser (CVPR 2019)	116.8	97.9	56.4	87.1	73.7	45.7
SD + ITW + BL	TokenHMR	**109.4**	**91.7**	**55.6**	**84.6**	**71.0**	**44.3**

Qualitative Results

Input · TokenHMR · HMR2.0

Resources

Project Page
https://tokenhmr.is.tue.mpg.de

Code
https://github.com/saidwivedi/TokenHMR