Action-Conditioned 3D Human Motion Synthesis with Transformer VAE

Mathis Petrovich[1] Michael J. Black[2] Gül Varol[1]

[1]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France

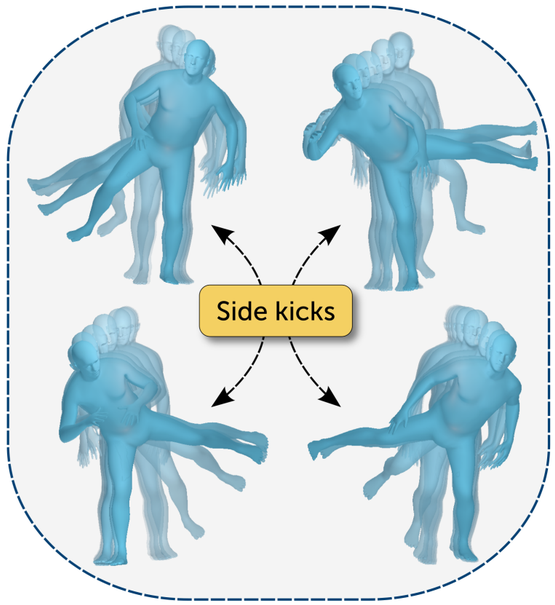
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

ICCV 2021 OCTOBER 11-17 VIRTUAL

https://imagine.enpc.fr/~petrovim/actor/

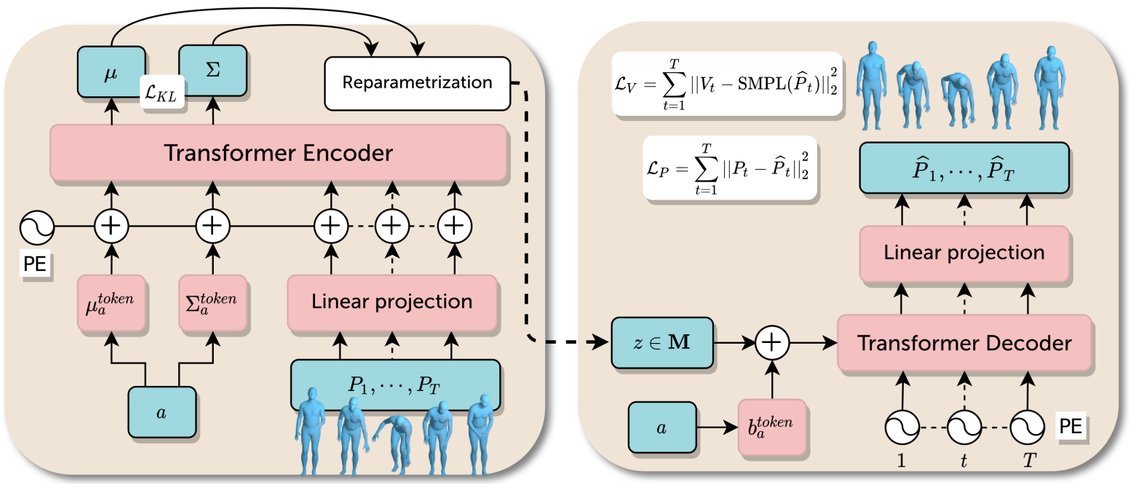
Introduction

- Goal & contributions
  - Generating synthetic but realistic and diverse human motion sequence given an action label
  - Learning from noisy 3D body poses estimated from monocular action recognition datasets[3][4][5]
- Motivations
  - Augmenting existing Mocap datasets, which are expensive and limited in size
  - Serving as additional training data for motion recognition
  - A compact action-aware latent space for human motions



ACTOR: Action-Conditioned TransfORmer VAE

- Non-autogressive Transformer architecture
- Sequence-level Variational autoencoder (VAE)
- Learnable tokens μ_a^{token} and Σ_a^{token}
- Loss terms on rotations and vertices (SMPL[1])
- Allows to generate variable length sequences with various body shapes

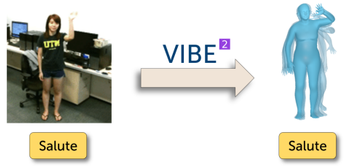


Encoder **Decoder**

Training data

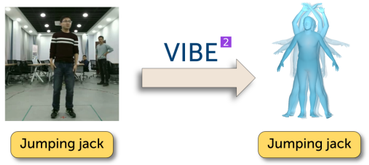
NTU13[4] (13 actions)

- RGB-D dataset, subset of NTU-120
- SMPL poses estimated with VIBE[2]



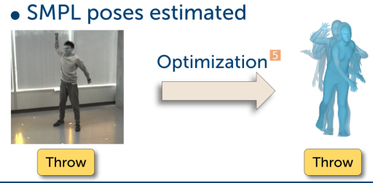
UESTC[3] (40 actions)

- RGB-D dataset
- SMPL poses estimated with VIBE[2]



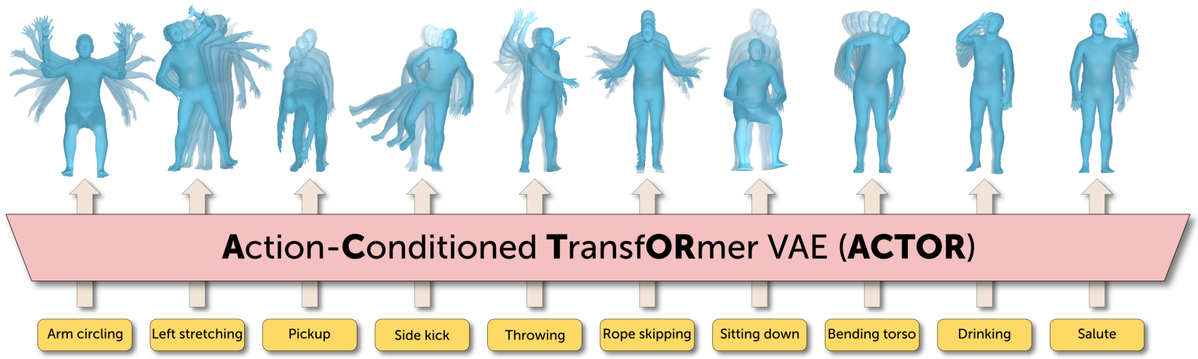
HumanAct12[5] (12 actions)

- RGB-D + polarization images, subset of PHSPDataset
- SMPL poses estimated



Qualitative results

The generated sequences are realistic, diverse and smooth.



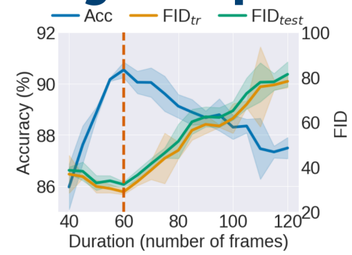
Generating variable-length sequences

We evaluate the capability of the models trained on UESTC[3] with

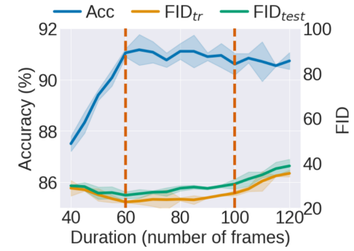
- fixed size (60 frames)
- variable-size (between 60/100 frames)

on generating various durations.

The performance is overall improved when the model has previously seen duration variations in training.



Trained with fixed duration



Finetuned with variable duration

Comparison with previous works

Metrics

FID: Fréchet Inception Distance (Similarity between GT distribution and the generation distribution)

Acc: Action recognition accuracy

Div: Diversity

Multimod: Multimodality (Per-action diversity)

	NTU-13[4]				HumanAct12[5]			
Method	FID_{tr} ↓	Acc. ↑	Div.→	Multimod.→	FID_{tr} ↓	Acc. ↑	Div.→	Multimod.→
Real [Action2Motion]	0.03	99.9	7.11	2.19	0.09	99.7	6.85	2.45
Real*	0.02	99.8	7.07	2.25	0.02	99.4	6.86	2.60
CondGRU	28.31	7.80	3.66	3.58	40.61	8.0	2.38	2.34
Two-stage GAN	13.86	20.2	5.33	3.49	10.48	42.1	5.96	2.81
Act-MoCoGAN	2.72	**99.7**	6.92	0.91	5.61	79.3	6.75	1.06
Action2Motion[5]	0.33	94.9	7.07	2.05	2.46	92.3	7.03	2.87
ACTOR (ours)	**0.11**	97.1	7.08	2.08	**0.12**	**95.5**	6.84	2.53

References

[1] Loper et al. SMPL: A skinned multi-person linear model 2015

[2] Kocabas et al VIBE: Video inference for human body pose and shape estimation 2020

[3] Ji et al. A large-scale RGB-D database for arbitrary-view human action recognition 2018

[4] Liu et al. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding 2019

[5] Guo et al. Action2Motion: Conditioned generation of 3D human motions 2020