Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation

Wenhao Li[1] Mengyuan Liu[1] Hong Liu[1] Pichao Wang[2] Jialun Cai[1] Nicu Sebe[3]

[1]Peking University [2]Amazon Prime Video [3]University of Trento

https://github.com/NationalGAILab/HoT

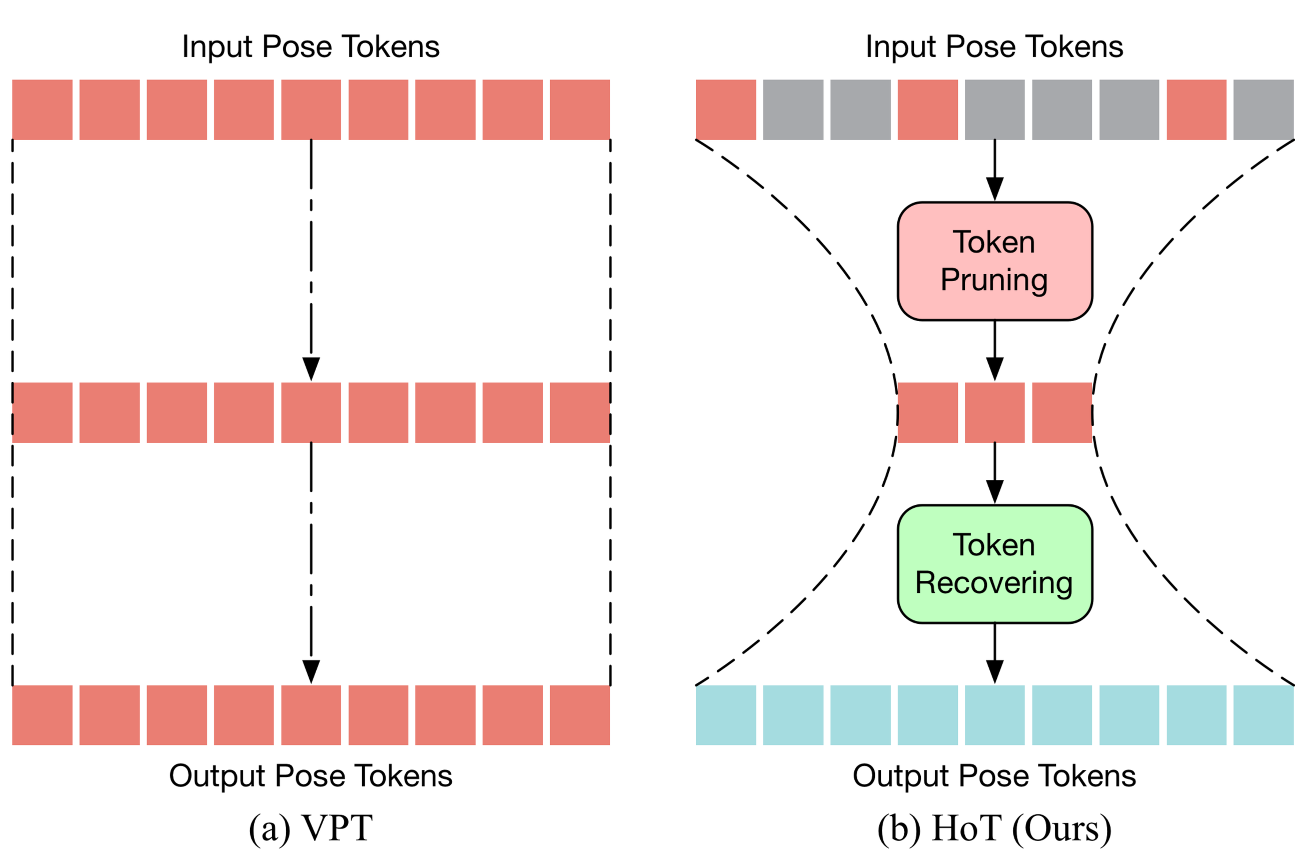
CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction

Transformers have been successfully applied in video-based 3D human pose estimation (HPE). However, the high computational costs of these video pose transformers (VPTs) make them impractical on resource-constrained devices.

How to achieve efficient VPTs?

- **Large receptive field**: Directly reducing the frame number can boost VPTs' efficiency, but it results in a small temporal receptive field that limits the model to capture richer spatio-temporal information to improve performance.
- **Video Redundancy**: Adjacent frames in a video sequence contain redundant information due to the similarity of nearby poses. Moreover, recent studies found that some tokens tend to be similar in the deep transformer blocks.
- ***Seq2seq* Inference**: For fast inference, A real-world 3D HPE system should be able to estimate the consecutive 3D poses of all frames at once in an input video.

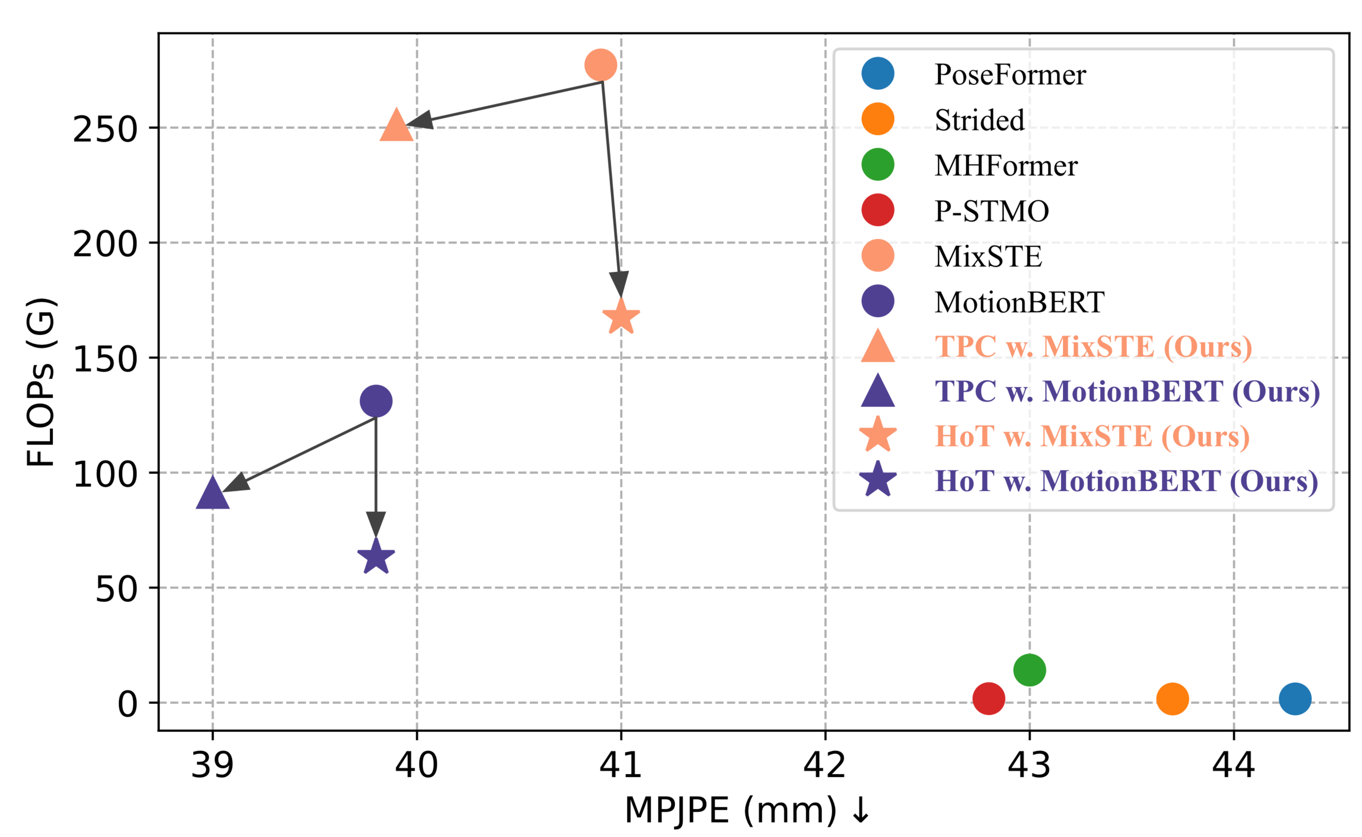


✨ ***Simple baseline, general-purpose efficient transformer-based framework***

🔥HoT🔥 is the first plug-and-play framework for efficient transformer-based 3D HPE. Unlike existing VPTs, which follow a "rectangle" paradigm that maintains the full-length sequence across all blocks, HoT begins with pruning the pose tokens of redundant frames and ends with recovering the full-length tokens (look like an "hourglass" ⏳). It is a general-purpose pruning-and-recovering framework, capable of being easily incorporated into common VPT models on both *seq2seq* and *seq2frame* pipelines while effectively accommodating various token pruning and recovery strategies.

Comparison with SOTA VPTs

Method	F	Param	FLOPs	MPJPE ↓
PoseFormer (ICCV'21) [51]	81	9.60	1.63	44.3
Strided (TMM'22) [17]	351	4.35	1.60	43.7
P-STMO (ECCV'22) [34]	243	7.01	1.74	42.8
STCFormer (CVPR'23) [35]	243	18.93	156.22	40.5
MHFormer (CVPR'22) [18]	351	31.52	14.15	43.0
TPC w. MHFormer (Ours)	351	31.52	8.22 (↓ **41.91%**)	43.0
MixSTE (CVPR'22) [48]	243	33.78	277.25	40.9
HoT w. MixSTE (Ours)	243	35.00	167.52 (↓ **39.6%**)	41.0
TPC w. MixSTE (Ours)	243	33.78	251.29 (↓ **09.4%**)	39.9
MotionBERT (ICCV'23) [52]	243	16.00	131.09	39.2
MotionBERT (ICCV'23) [52]*	243	16.00	131.09	39.8
HoT w. MotionBERT (Ours)	243	16.35	63.21 (↓ **51.8%**)	39.8
TPC w. MotionBERT (Ours)	243	16.00	91.38 (↓ **30.3%**)	**39.0**

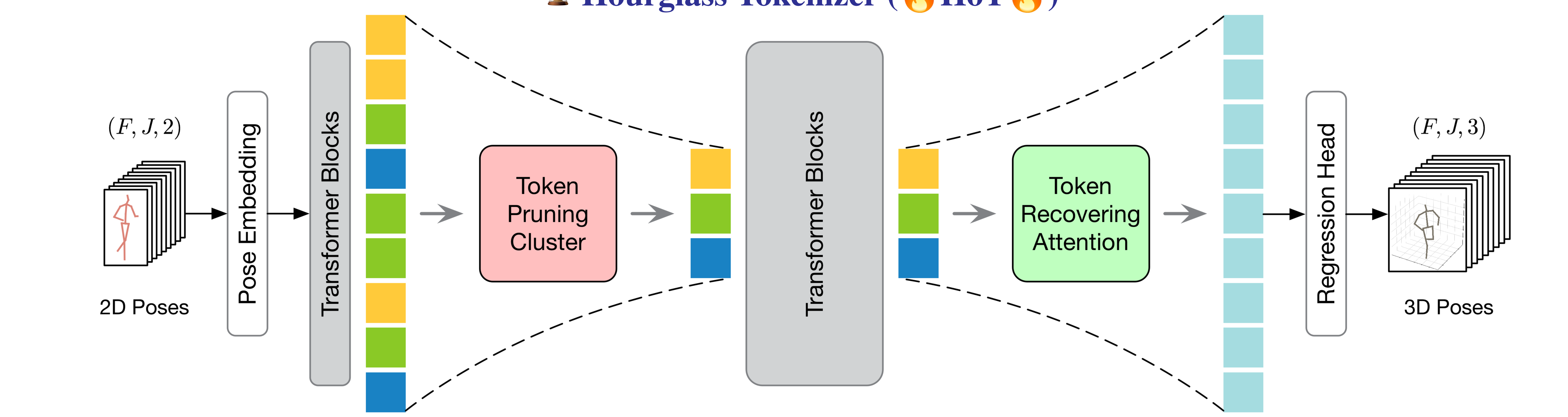


🚀 ***Both high efficiency and estimation accuracy***

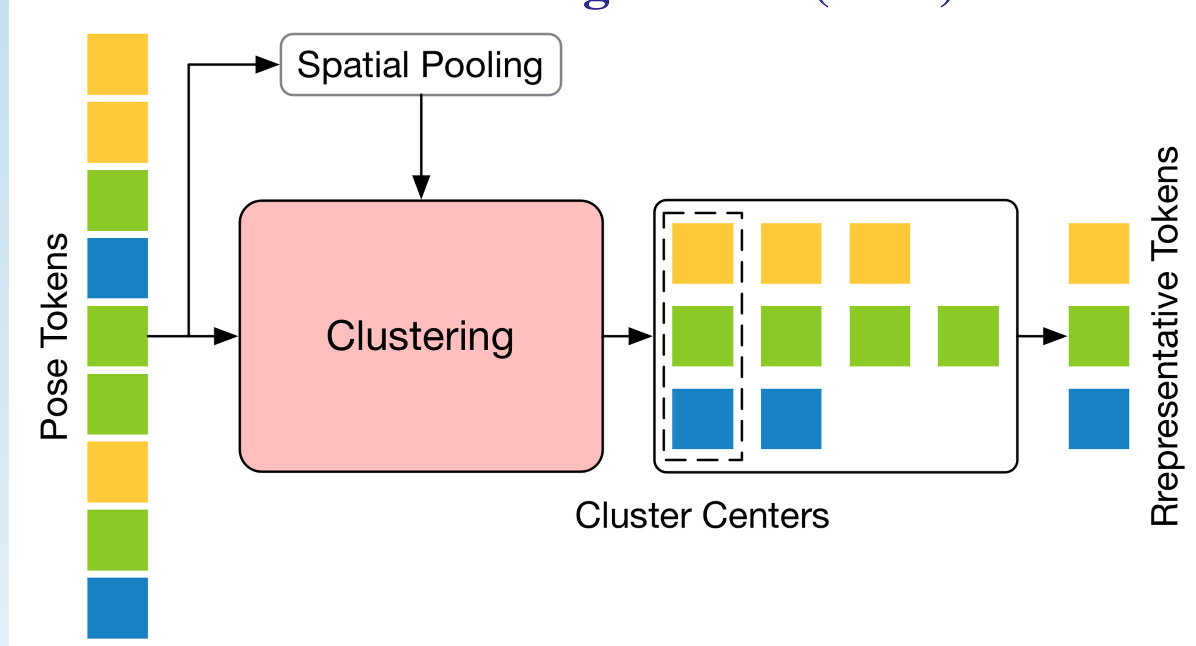
🔥HoT🔥 reveals that maintaining the full-length pose sequence is redundant, and a few pose tokens of representative frames can achieve both high efficiency and performance. Our HoT achieves highly competitive or even better results while bringing significant improvements in efficiency compared to the original VPTs.

Method

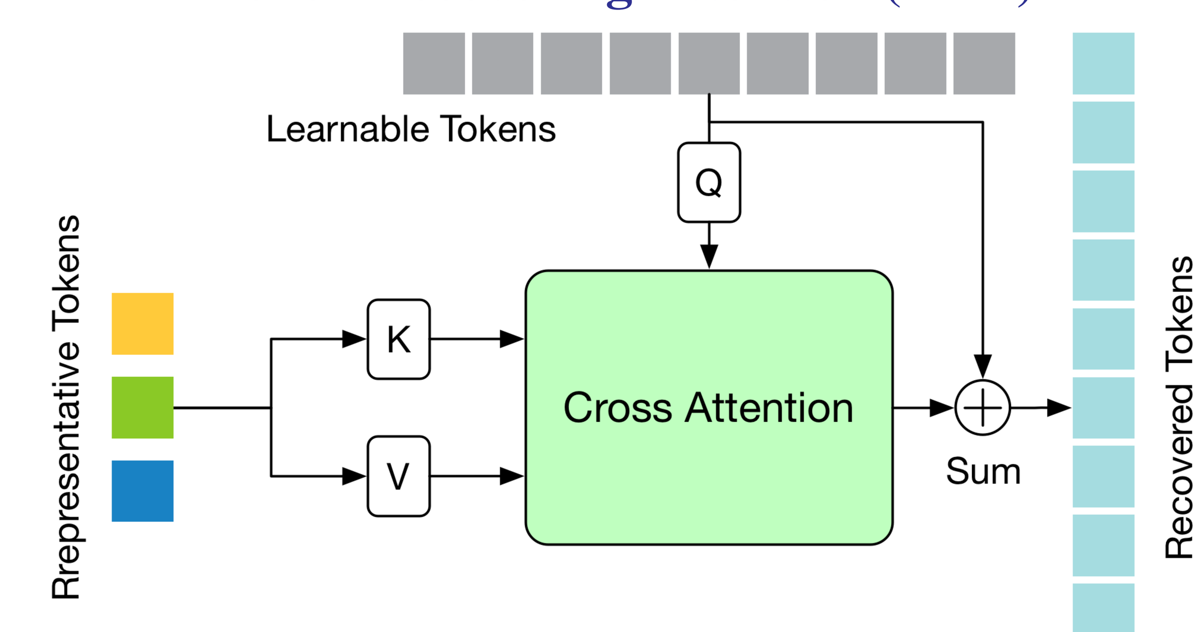
⏳ Hourglass Tokenizer (🔥HoT🔥)



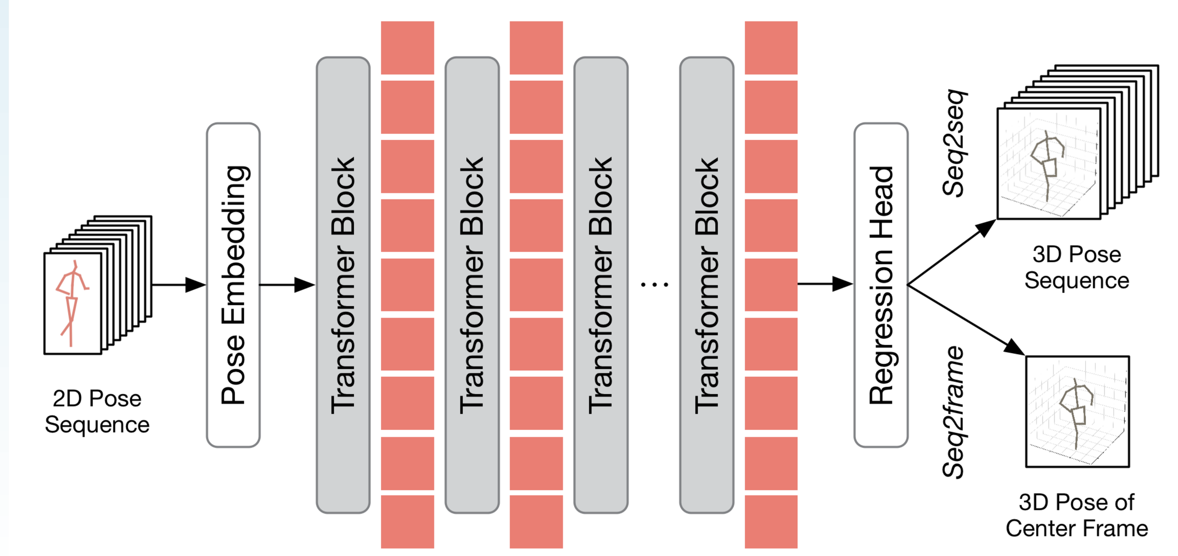
Token Pruning Cluster (TPC)



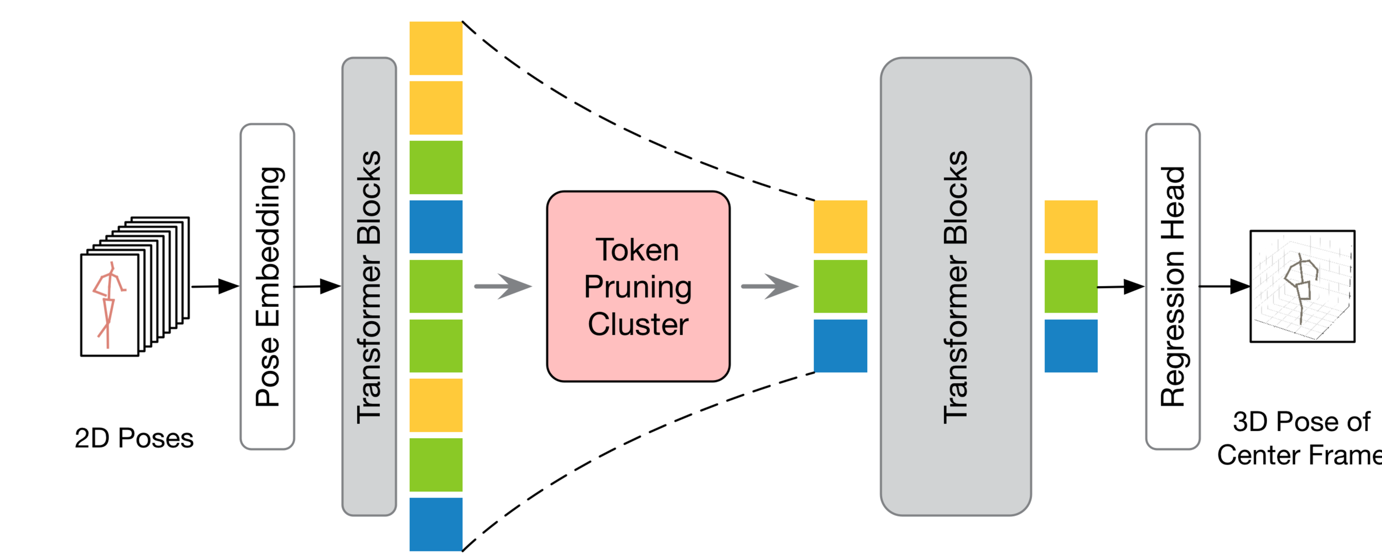
Token Recovering Attention (TRA)



Summary of VPT architecture



Our framework on *seq2frame* pipeline



Ablation Study

Ablation study on *seq2seq* (∗) and *seq2frame* (†)

Method	Param (M)	FLOPs (G)	FPS	MPJPE ↓
MixSTE [48] (∗)	33.78	277.25	10432	40.9
HoT w. MixSTE (∗)	35.00	167.52 (↓ **39.6%**)	15770 (↑ **51.2%**)	41.0
MixSTE [48] (†)	33.78	277.25	43	40.7
TPC w. MixSTE (†)	33.78	161.73 (↓ **41.7%**)	68 (↑ **58.1%**)	40.4
MotionBERT [52] (∗)	16.00	131.09	14638	39.8
HoT w. MotionBERT (∗)	16.35	63.21 (↓ **51.8%**)	25526 (↑ **74.4%**)	39.8
MotionBERT [52] (†)	16.00	131.09	60	39.5
TPC w. MotionBERT (†)	16.00	61.04 (↓ **53.4%**)	109 (↑ **81.7%**)	39.2

Ablation study on token pruning and recovering

		seq2seq			*seq2frame*	
Method	FN	Full ↓	Pruned ↓	Selected ↓	Center ↓	Selected ↓
MixSTE [48]	6.61	40.9	-	-	40.7	-
Ours, Uniform Sampling	6.61	41.4	41.3	41.4	40.7	40.8
Ours, Attention Pruning	6.56	42.1	42.5	41.5	42.3	44.4
Ours, Motion Pruning	7.00	42.8	43.4	41.6	41.3	42.3
Ours, the Proposed TPC	6.63	**41.0**	**41.3**	**40.2**	**40.4**	**39.4**

Method	Param	FLOPs	Full ↓	Pruned ↓	Selected ↓	Δ
MixSTE [48]	33.78	277.25	40.9	-	-	-
Ours, Nearest Interpolation	33.78	161.73	41.5	42.2	40.2	2.0
Ours, Linear Interpolation	33.78	161.73	41.3	41.9	**40.0**	1.9
Ours, the Proposed TRA	35.00	167.52	**41.0**	**41.3**	40.2	**1.1**