MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video

Jinlu Zhang[1] Zhigang Tu[1] Jianyu Yang[2] Yujin Chen[3] Junsong Yuan[4]

[1]*Wuhan University*
[2]*Soochow University*
[3]*Technical University of Munich*
[4]*State University of New York at Buffalo*

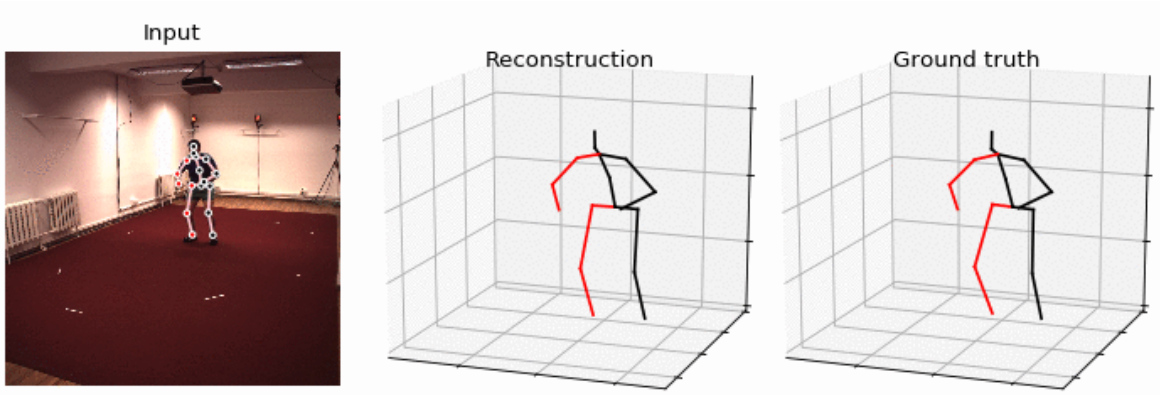
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Introduction of 3D Human Pose Estimation

- **Definition**
  - 3D human pose estimation from monocular observations is a fundamental vision task that reconstructs 3D body joint locations from the input images or video.

- **Challenge**
  - Depth Ambiguity
  - Temporal Consistency
  - 2D Input Uncertainty
  - Occlusion
  - ……

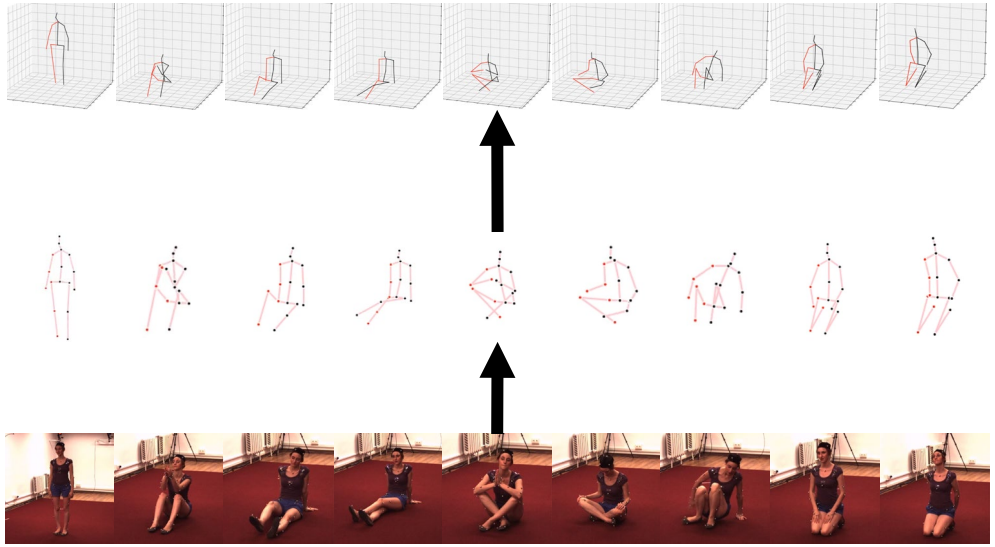
Learning better spatio-temporal correlation
is very important to address these challenges



Input：2D Pose Sequence

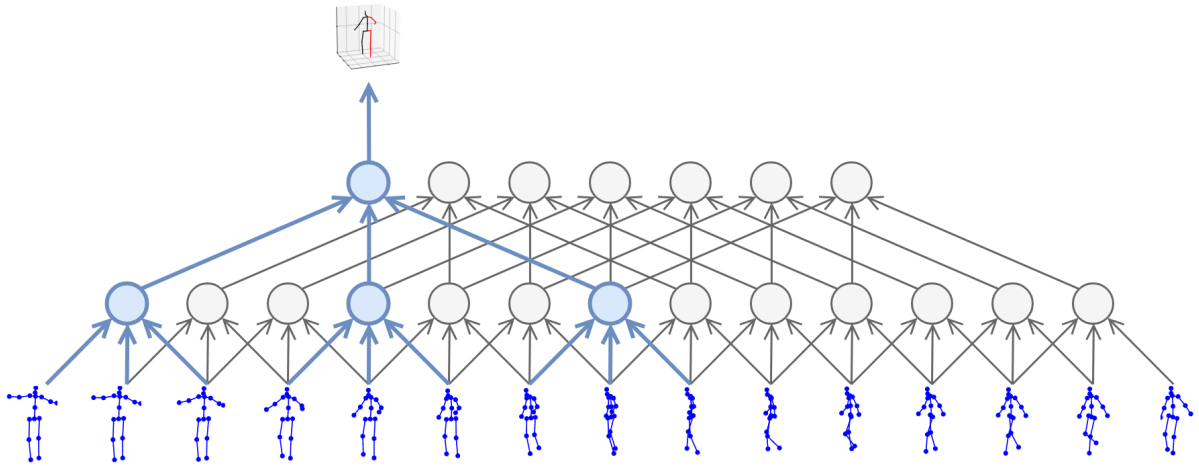
Output：Root-Relative 3D coordinates

Related Works of 3D Pose Estimation in Video



Seq2seq paradigm[1]

1. **Seq2seq** methods based on LSTM/RNN can not parallel over the time.
2. Most of them lack the global modeling ability on **long sequences.**



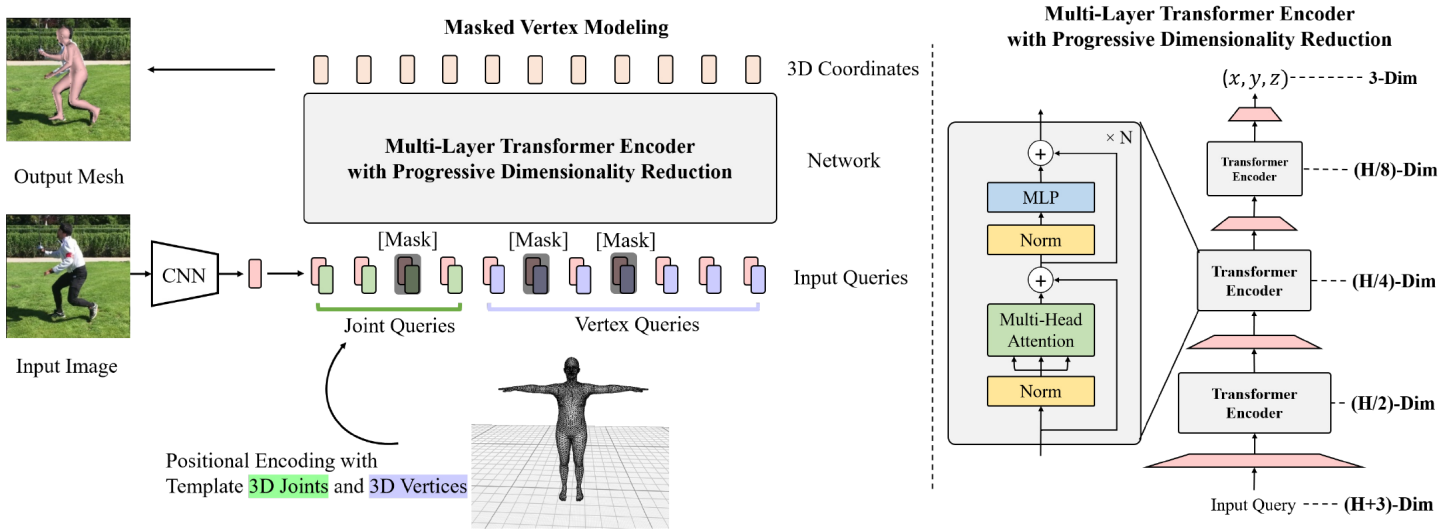
Seq2frame paradigm[2]

Spatio-Temporal correlation modeling is insufficient and low-efficiency

1. Seq2frame methods ignore the **temporal motion** among body joints;
2. Most SOTA methods increase the **redundant calculation**.

[1] K. Lee et al. Propagating LSTM: 3D Pose Estimation based on Joint Interdependency, 2018, ECCV
[2] Pavllo et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training, 2019, CVPR

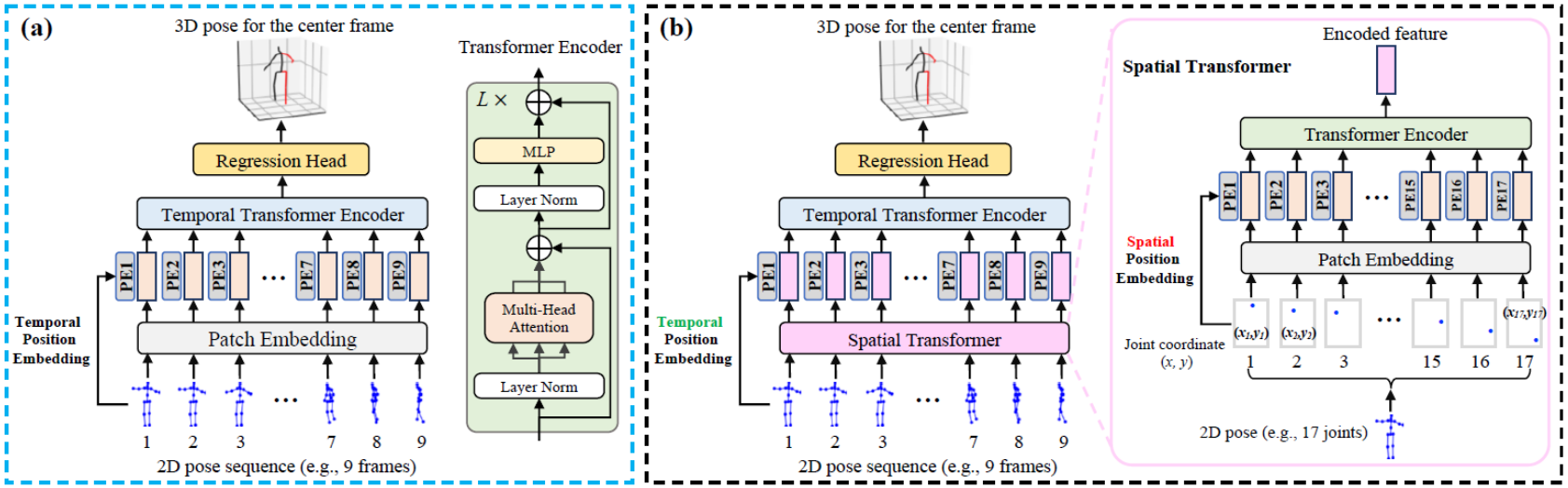
Recent Works of Transformer in 3D Pose Estimation



Single-frame method

Lin et al. End-to-End Human Pose and Mesh Reconstruction with Transformers, 2021, CVPR

1. Jitter results **in video input.**
2. Single frame performance is still **not good enough**.

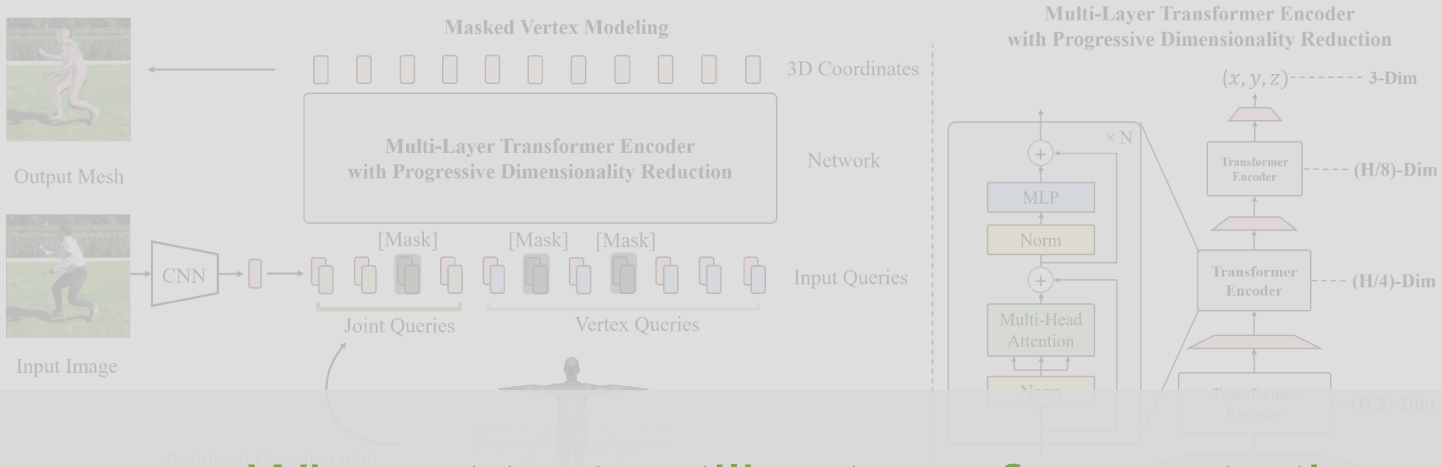


Video method (PoseFormer)

Zheng et al. 3d human pose estimation with spatial and temporal transformers, 2021, ICCV

1. Much **redundant computation** during training and inference.
2. Poor performance **under seq2seq setting**, even transformer is a sequential model.

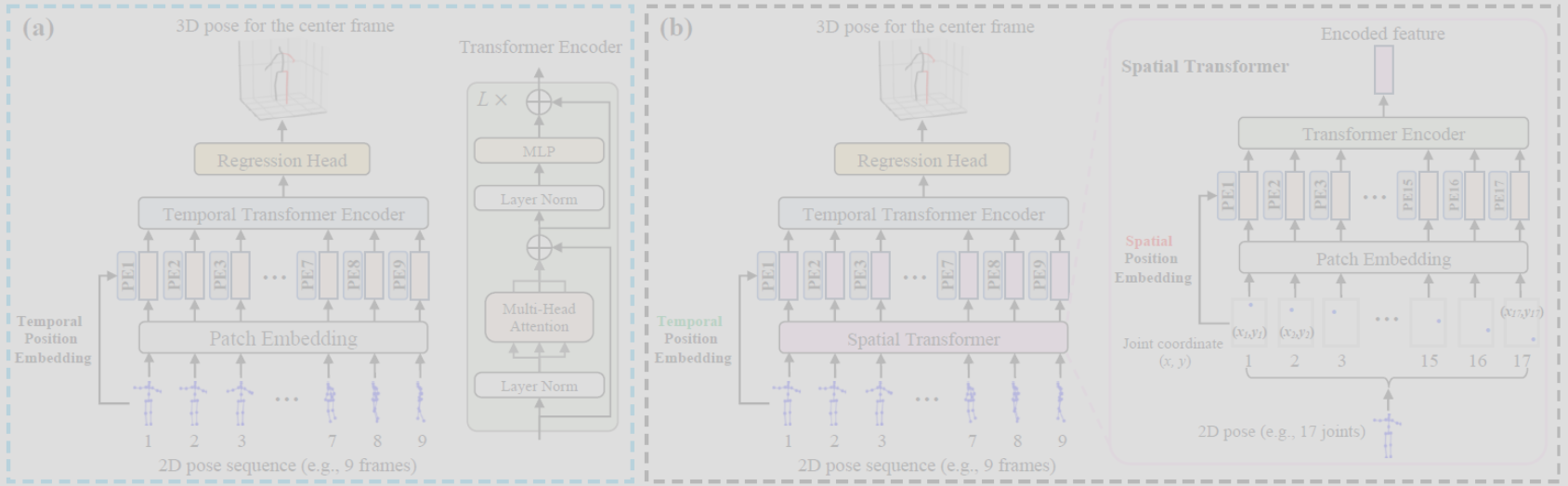
Recent Works of Transformer in 3D Pose Estimation



1. Jitter results **in video input.**
2. Single frame performance is still **not good enough**.

Why not try to utilize **transformer** to the **seq2seq paradigm** to obtain high efficiency and good performance?

Lin et al. End-to-End Human Pose and Mesh Reconstruction with Transformers, 2021, CVPR



1. Much **redundant computation** during training and inference.
2. Poor performance **under seq2seq setting**, even transformer is a sequential model.

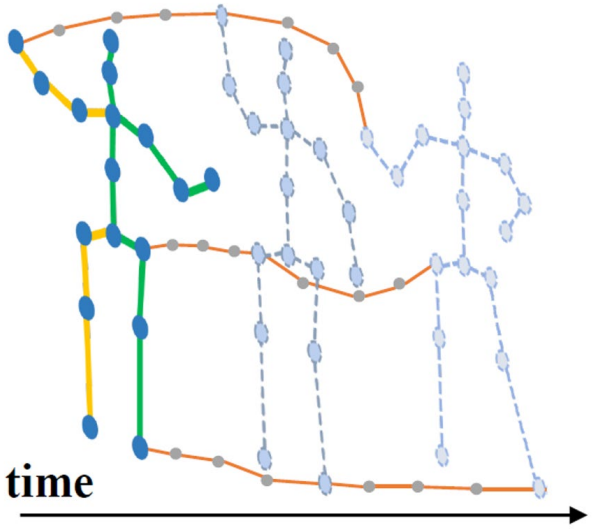
Video method (PoseFormer)

Zheng et al. 3d human pose estimation with spatial and temporal transformers, 2021, ICCV

Challenges

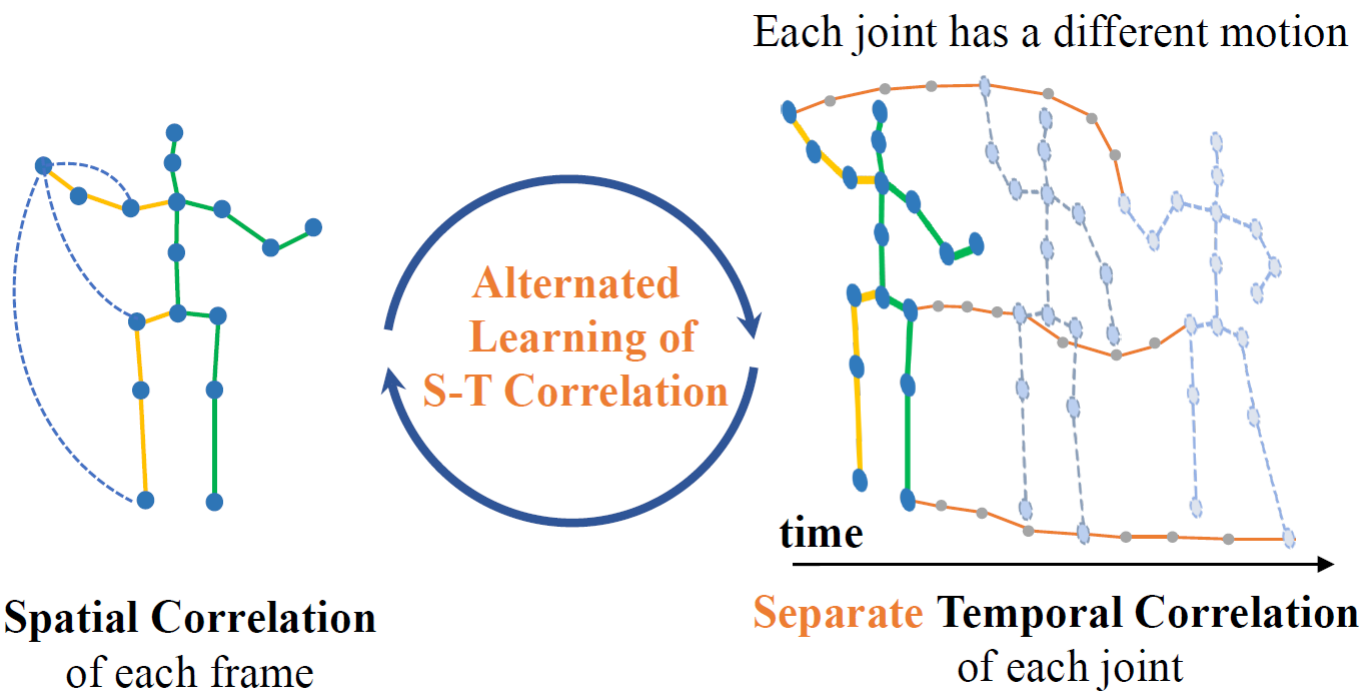
- How to design a model to obtain **better spatio-temporal modeling** ability?
- How to take advantage of global **correlations between sequences in video**?

Our Motivation



Each joint has different motion in a video sequence

Our Method

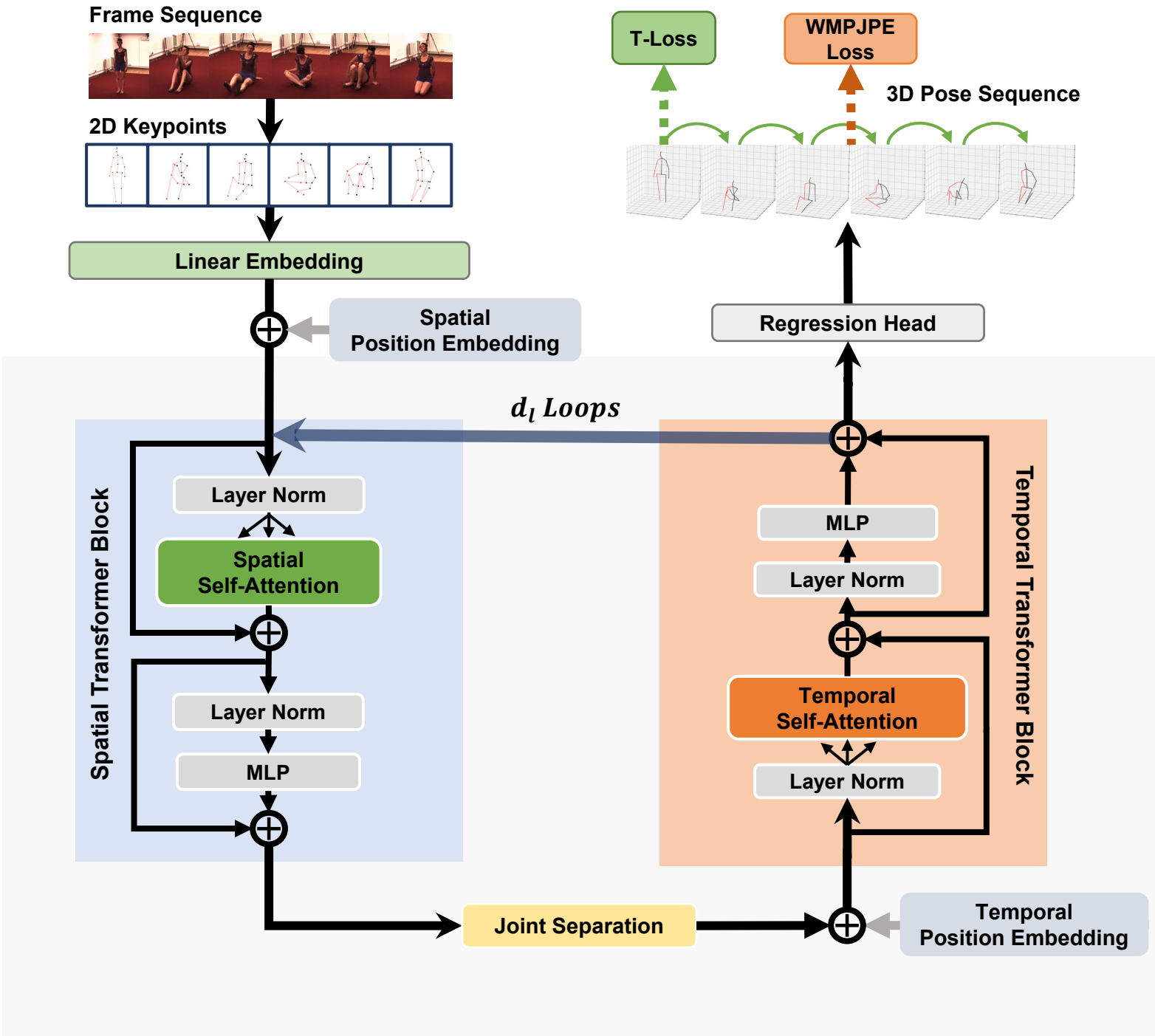


1. **Joint Separation**
   We separate different joints in time dimension, so that the trajectory of each joint is an individual token, and different joints of body are modeled paralleled.

2. **Alternating Design**
   We stack spatial and temporal encoders for d_l loops, and the dimension of feature is preserved as a fixed size d_m to promise that spatial-temporal correlation learning focuses on the same joint.

Our Method — MixSTE

(Mixed Spatial-Temporal Encoder)



Main Experimental Results

Comparison results under Protocol 1 (MPJPE) on Human3.6M using detectors

Protocol #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg.**
Pavlakos *et al.* [35]	CVPR2018	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
Pavllo *et al.* [37](CPN, T=243)(†)	CVPR2019	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Cai *et al.* [1](CPN, T=7)(†)	ICCV2019	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Yeh *et al.* [51](†)	NIPS2019	44.8	46.1	43.3	46.4	49.0	55.2	44.6	44.0	58.3	62.7	47.1	43.9	48.6	32.7	33.3	46.7
Liu *et al.* [28](CPN, T=243)(†)	CVPR2020	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Wang *et al.* [46](CPN, T=96)(†)	ECCV2020	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
Chen *et al.* [4](CPN, T=243)(†)	TCSVT2021	41.4	43.5	40.1	42.9	46.6	51.9	41.7	42.3	53.9	60.2	45.4	41.7	46.0	31.5	32.7	44.1
Xu *et al.* [48](T=1)	CVPR2021	45.2	49.9	47.5	50.9	54.9	66.1	48.5	46.3	59.7	71.5	51.4	48.6	53.9	39.9	44.1	51.9
Lin *et al.* [25](T=1)(*)	CVPR2021	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	54.0
Zeng *et al.* [53](†)	ICCV2021	43.1	50.4	43.9	45.3	46.1	57.0	46.3	47.6	56.3	61.5	47.7	47.4	53.5	35.4	37.3	47.9
Zheng *et al.* [57](CPN, T=81)(†)(*)	ICCV2021	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Ours(CPN, T=81)(†)(*)		39.8	43.0	38.6	40.1	43.4	50.6	40.6	41.4	52.2	56.7	43.8	40.8	43.9	29.4	30.3	42.4
Ours(CPN, T=243)(†)(*)		**37.6**	**40.9**	**37.3**	**39.7**	**42.3**	**49.9**	**40.1**	**39.8**	**51.7**	**55.0**	**42.1**	**39.8**	**41.0**	**27.9**	**27.9**	**40.9**
Wang *et al.* [46](HRNet, T=96)(†)	ECCV2020	38.2	41.0	45.9	39.7	41.4	51.4	41.6	41.4	52.0	57.4	41.8	44.4	41.6	33.1	**30.0**	42.6
Wehrbein *et al.* [47](HRNet, T=200)	ICCV2021	38.5	42.5	39.9	41.7	46.5	51.6	39.9	40.8	49.5	56.8	45.3	46.4	46.8	37.8	40.4	44.3
Ours(HRNet, T=243)		**36.7**	**39.0**	**36.5**	**39.4**	**40.2**	**44.9**	**39.8**	**36.9**	**47.9**	**54.8**	**39.6**	**37.8**	**39.3**	**29.7**	30.6	**39.8**

Comparison results under Protocol 1 (MPJPE) on Human3.6M using 2D ground truth

Protocol #1		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	**Avg.**
Liu *et al.* [28](T=243)(†)	CVPR2020	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Wang *et al.* [46](GT, T=96)	ECCV2020	23.0	25.7	22.8	22.6	24.1	30.6	24.9	24.5	31.1	35.0	25.6	24.3	25.1	19.8	18.4	25.6
Zheng *et al.* [57](T = 81)(†)(*)	ICCV2021	30.0	33.6	29.9	31.0	30.2	33.3	34.8	31.4	37.8	38.6	31.7	31.5	29.0	23.3	23.1	31.3
Ours(T=81)		25.6	27.8	24.5	25.7	24.9	29.9	28.6	27.4	29.9	29.0	26.1	25.0	25.2	18.7	19.9	25.9
Ours(T=243)		**21.6**	**22.0**	**20.4**	**21.0**	**20.8**	**24.3**	**24.7**	**21.9**	**26.9**	**24.9**	**21.2**	**21.5**	**20.8**	**14.7**	**15.7**	**21.6**

MixSTE as a seq2seq method first outperforms the seq2frame methods!

Main Experimental Results

Comparison on MPI-INF-3DHP with 2D GT

Method		PCK↑	AUC↑	MPJPE↓
Mehta *et al.* [28]	ACM TOG 2017	79.4	41.6	-
Lin *et al.* [20](T=25)	BMVC2019	83.6	51.4	79.8
Li *et al.* [18]	CVPR2020	81.2	46.1	99.7
Wang *et al.* [41](T=96)	ECCV2020	86.9	62.1	68.1
Chen *et al.* [4](T=243)	TCSVT2021	87.8	53.8	79.1
Gong *et al.* [8]	CVPR2021	88.6	57.3	73.0
Zheng *et al.* [50]	ICCV2021	88.6	56.4	77.1
Ours(T=1)		<u>94.2</u>	<u>63.8</u>	<u>57.9</u>
Ours(T=27)		**94.4**	**66.5**	**54.9**

Table 3. Detailed quantitative comparison results on MPI-INF-3DHP with three metrics. The ↑ indicates the higher, the better, the ↓ indicates the lower, the better. The best and second-best results are highlighted in bold and underlined formats, respectively.

Comparison on HumanEva-I with 2D GT

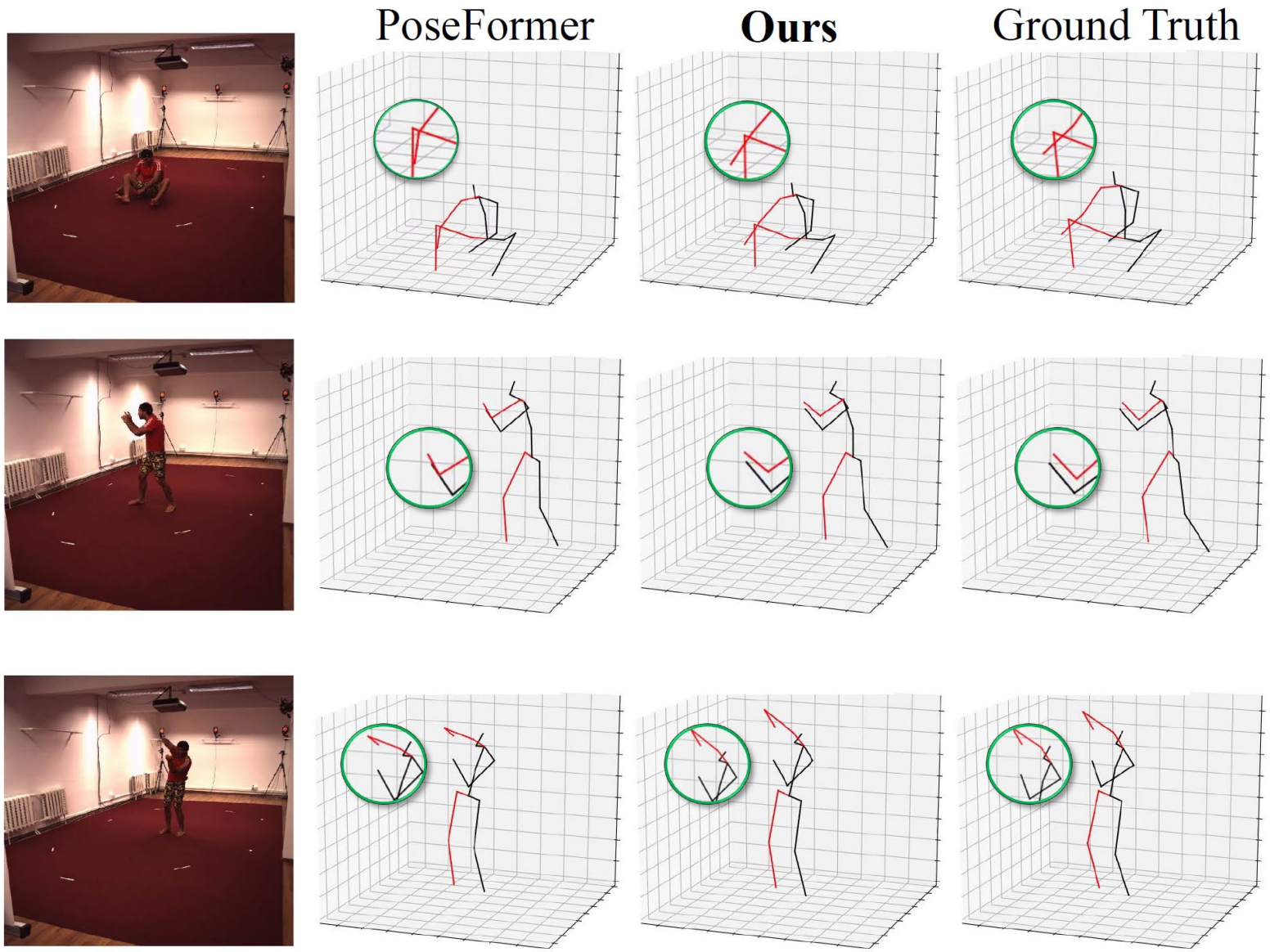
#Protocol1	Walk			Jog			Avg.
Pavllo *et al.* [32](T=81)	13.1	**10.1**	39.8	20.7	13.9	15.6	18.9
Pavllo *et al.* [32](T=81, FT)	14.0	12.5	27.1	**20.3**	17.9	17.5	18.2
Zheng *et al.* [50](T=43)	16.3	11	47.1	25	15.2	15.1	21.6
Zheng *et al.* [50](T=43, FT)	14.4	10.2	46.6	22.7	13.4	13.4	20.1
Ours(T=43)	20.3	22.4	34.8	27.3	32.1	34.3	28.5
Ours(T=43, FT)	**12.7**	10.9	**17.6**	22.6	**12.8**	**13.0**	**13.0**

Table 4. The MPJPE on HumanEva testset under Protocol 1. FT indicates using the pretrained model on Human3.6M for finetuning. The best result is highlighted in bold.

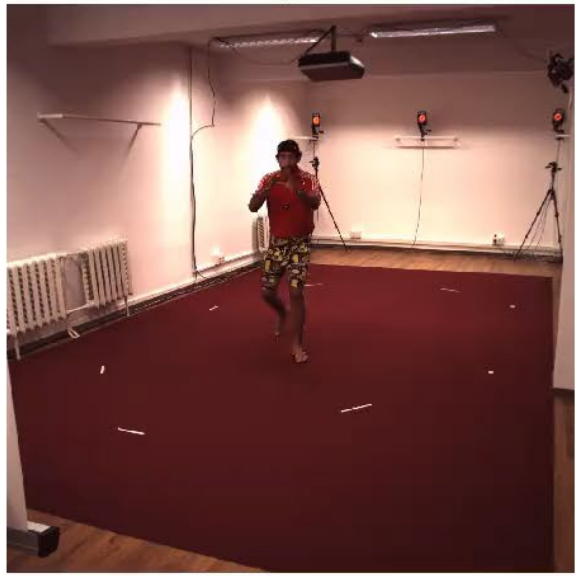
H36M SOTA Leaderboard

3DHP SOTA Leaderboard

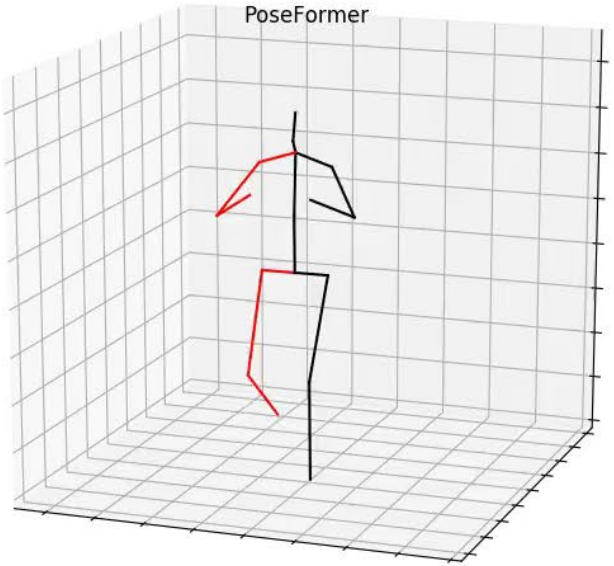
Qualitative Results Presentation


PoseFormer
Ours
Ground Truth

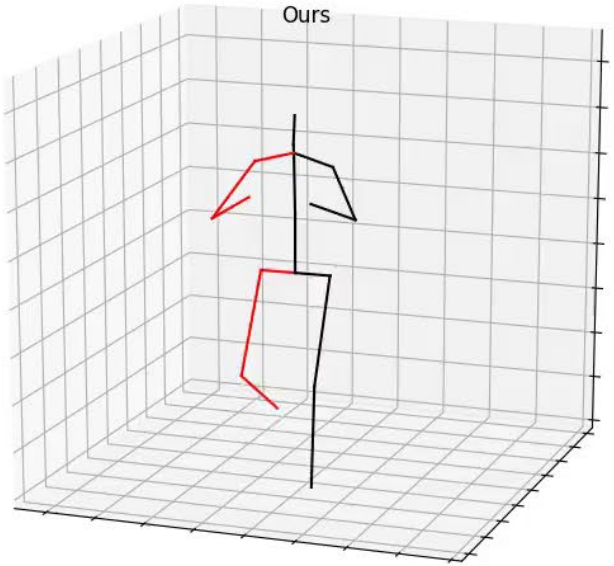
Input



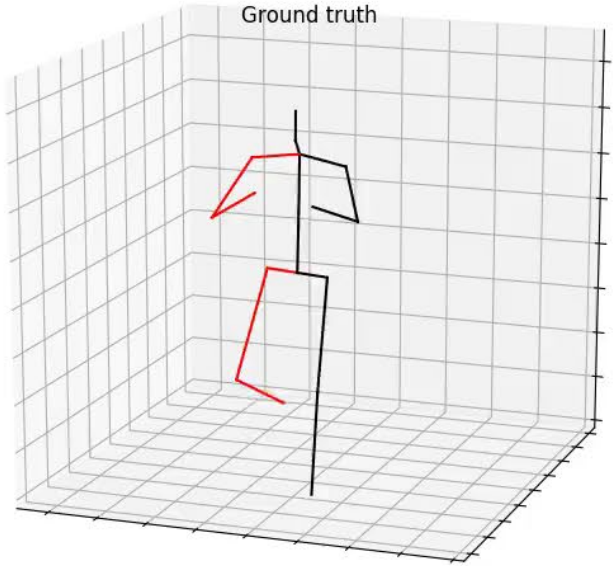
PoseFormer



Ours

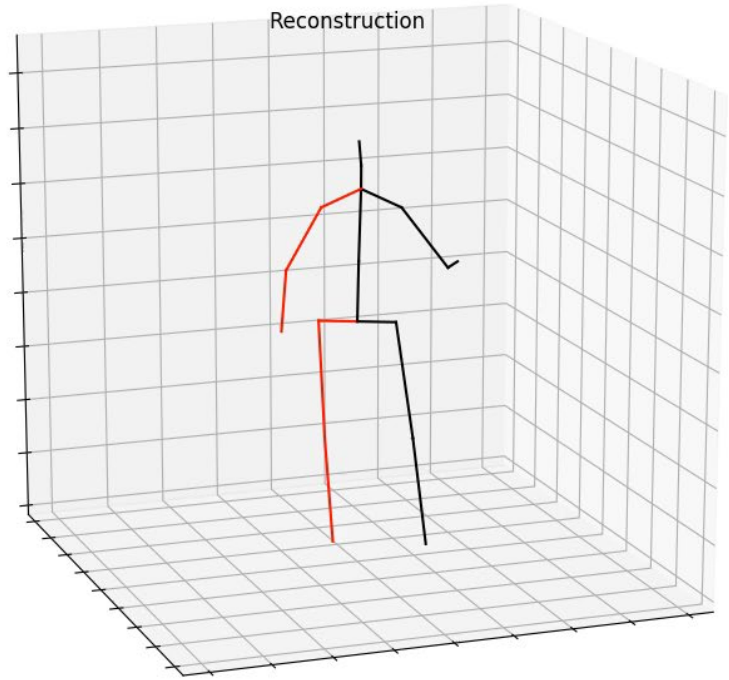


Ground truth



Input

Reconstruction



Thank you for watching!

HomePage: jinluzhang.site

Email: jinluzhang@whu.edu.cn

Code: https://github.com/JinluZhang1126/MixSTE

Code Link

HomePage