JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

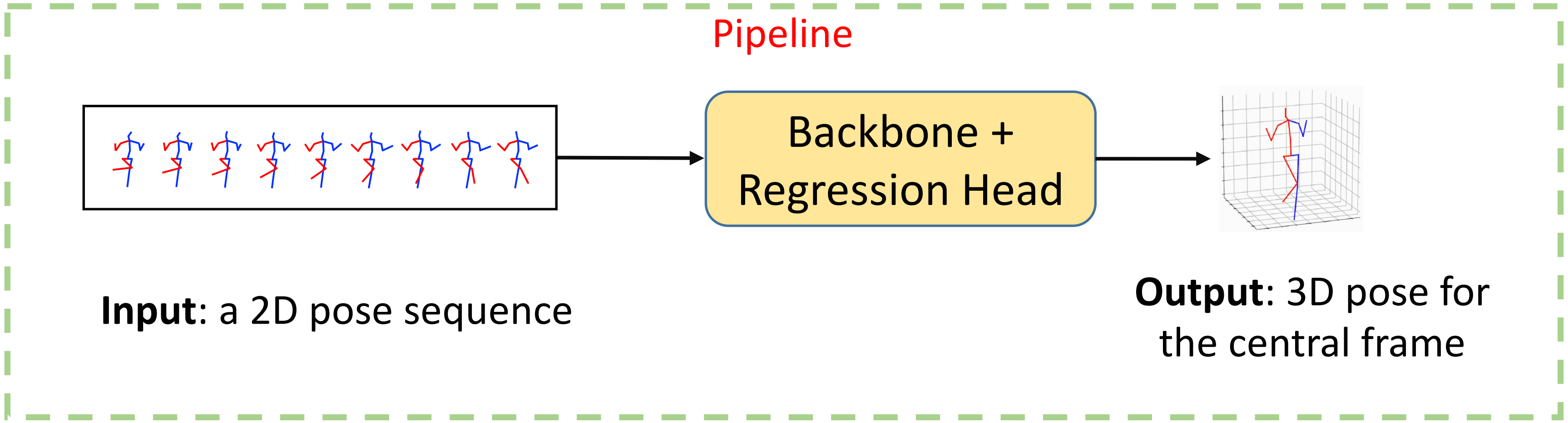
PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation

Poster: WED-AM-062

Qitao Zhao[1], Ce Zheng[2], Mengyuan Liu[3], Pichao Wang[4], Chen Chen[2]

[1]Shandong University,
[2]Center for Research in Computer Vision, University of Central Florida,
[3]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School,
[4]Amazon Prime Video

Introduction



Estimating the 3D human pose from a 2D pose sequence is now dominant in the literature (referred to as 2D-to-3D lifting methods).

Research Problems

State-of-the-art methods suffer from two **limitations**:

- **Poor efficiency** in temporal modeling for long joint sequences

  Applying dense temporal modeling (e.g., using self-attention) for **all video frames** is computationally expensive

- **Vulnerable** to the **noise** brought by imperfect 2D joint detection

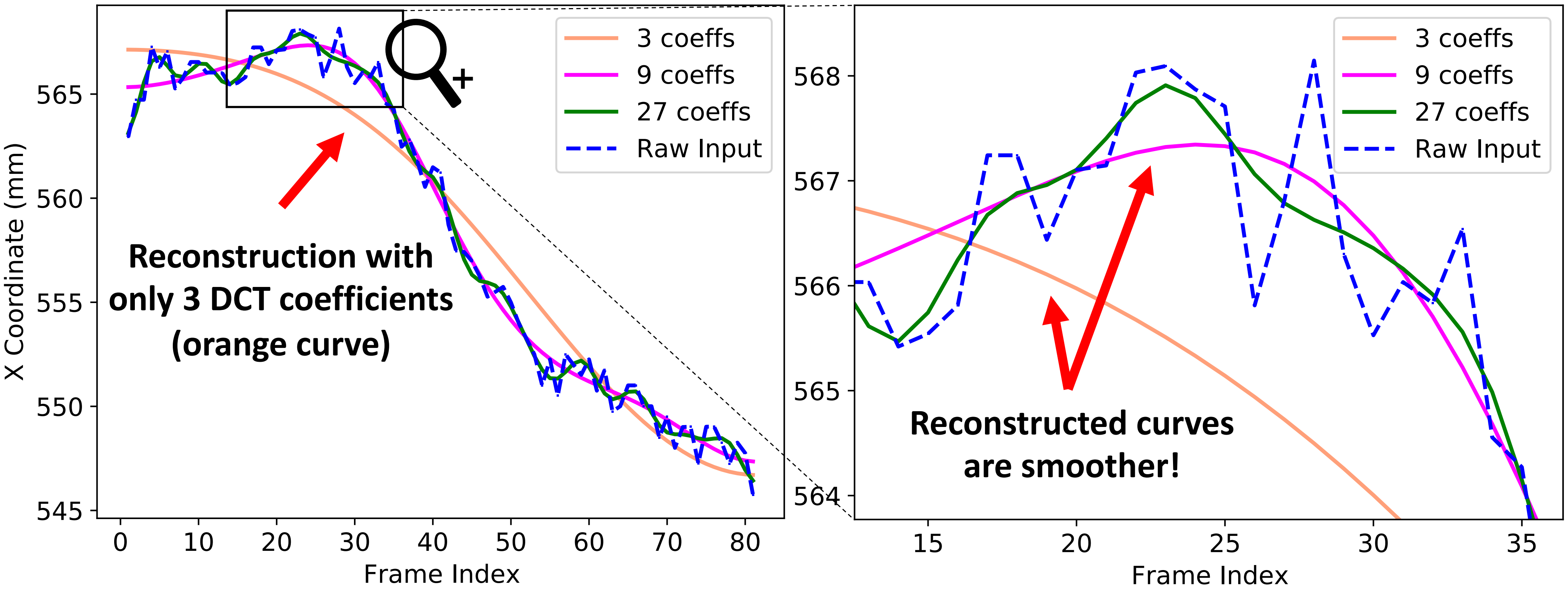
  Frame-to-frame interactions unexpectedly propagate and amplify the noise in each video frame

Take-away Message:

We find the **frequency-domain representation** of input sequences a surprising fit to **simultaneously** solve these two practical problems.

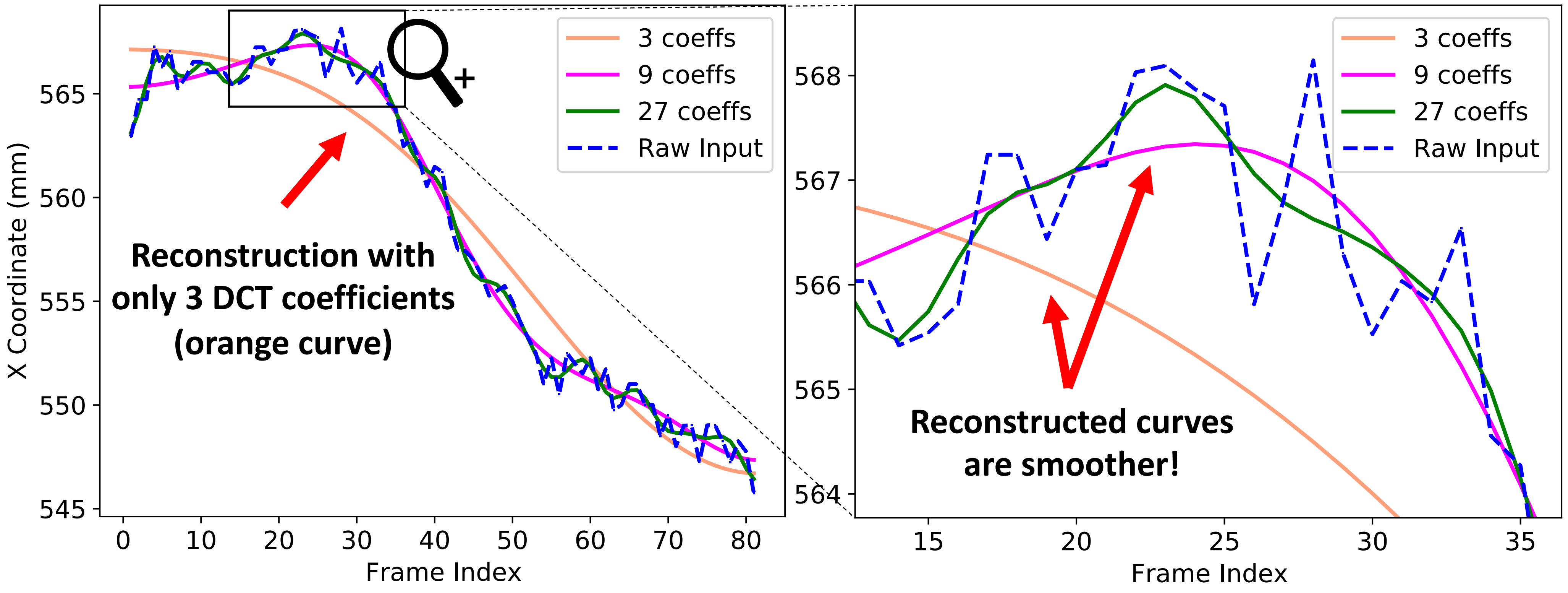
Motivation

We show a sample of joint trajectory and its reconstructions with a few low-frequency Discrete Cosine Transform (DCT) coefficients.



Motivation

Low-frequency coefficients
- are enough to encode **global human dynamics**
- **filter out noise** in the joint trajectory ("smoother")



The Proposed Method

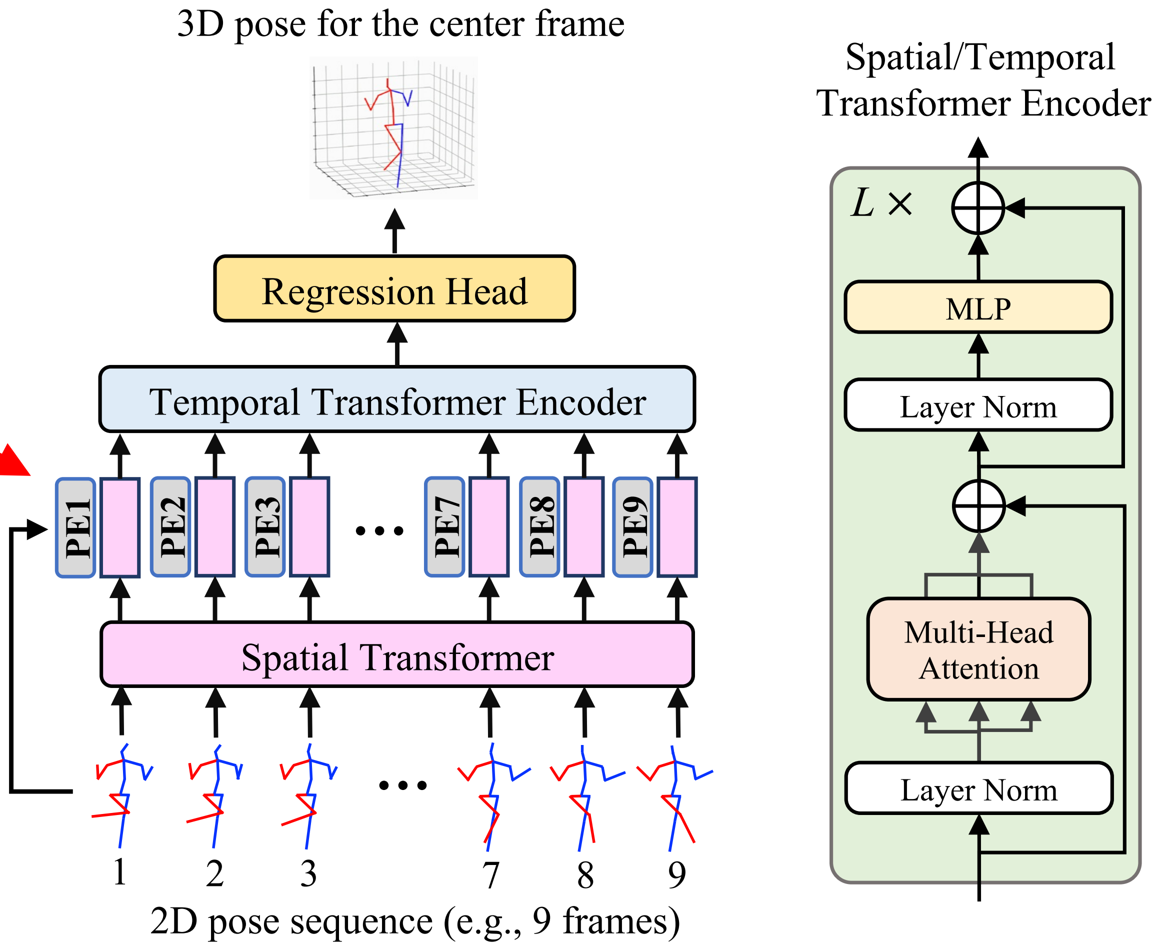
PoseFormer

We build our method on PoseFormer(V1).

PoseFormer

Here is an overview:

PoseFormer applies expensive self-attention to **all frames**

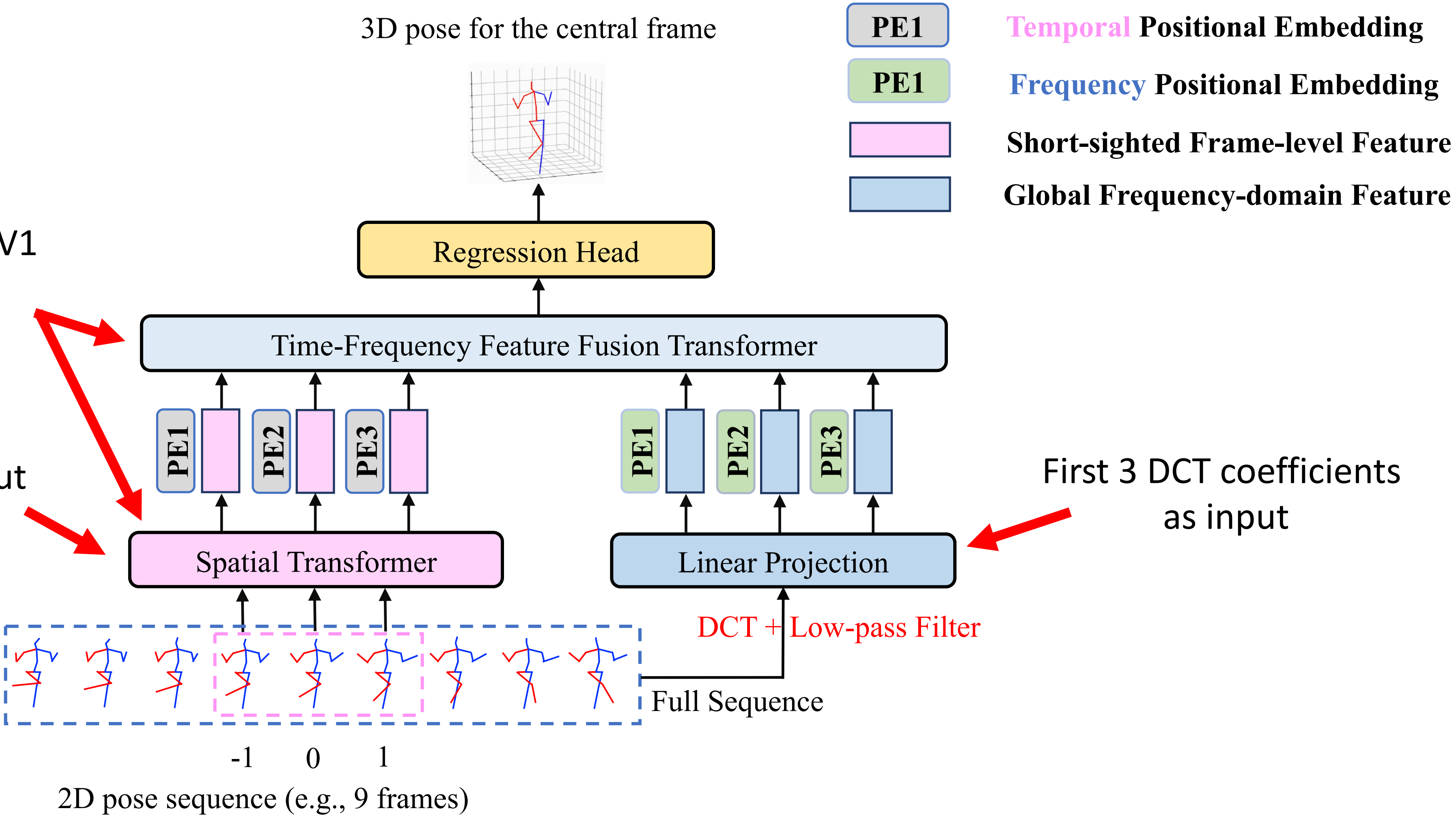


Zheng et al. 3d human pose estimation with spatial and temporal transformers. In ICCV, 2021.

The Proposed Method

Compared to PoseformerV1 the token number for transformers is reduced

3 central frames as input



First 3 DCT coefficients as input

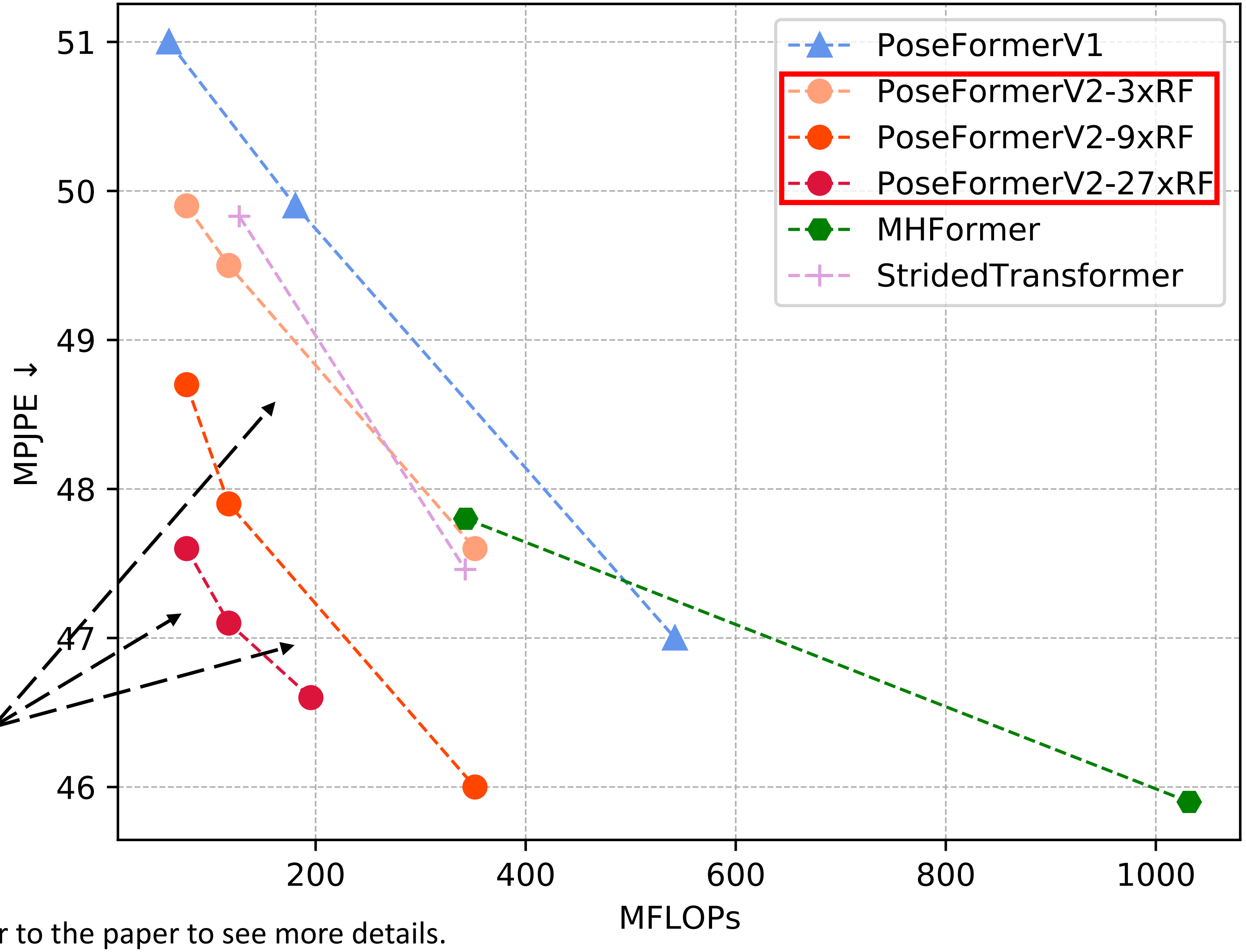
Zheng et al. 3d human pose estimation with spatial and temporal transformers. In ICCV, 2021.

Properties

- Given a long sequence, we only use a few central frames and its low-frequency coefficients, thus **reducing the effective sequence length**.
- The frame number and coefficient number can be **arbitrarily specified** for **a flexible speed-accuracy trade-off**.
- Low-frequency DCT coefficients **filter out noise** in the input 2D pose sequence and therefore **improve robustness**.

Comparisons with State-of-the-art Methods

Comparisons on Human3.6M



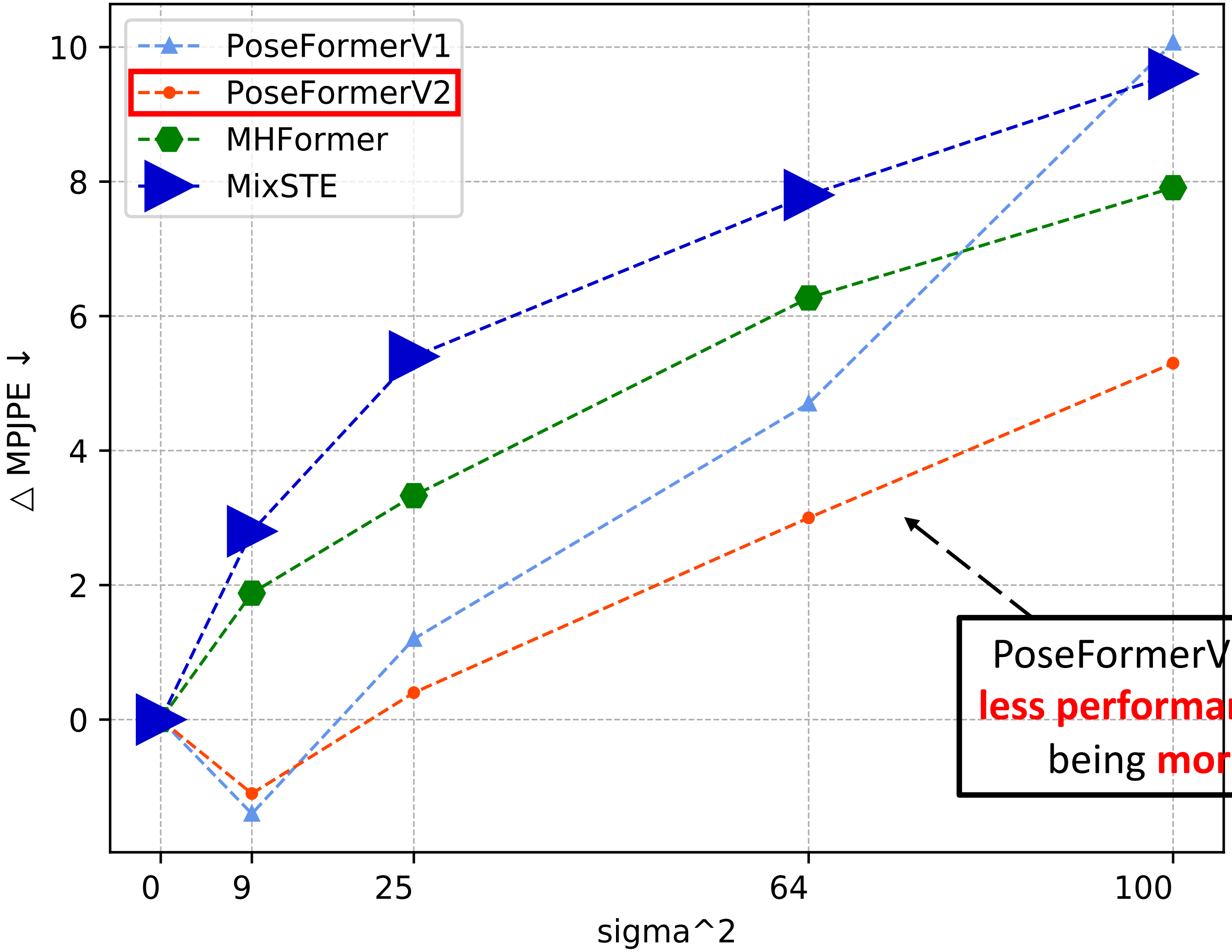
Superior speed-accuracy trade-off than other transformer-based methods

*RF denotes Receptive Field, please refer to the paper to see more details.

Comparisons on Human3.6M

We investigate the **robustness** of models by adding Gaussian noise to the ground-truth 2D joint detection of standard deviation *sigma*, and to show their **performance drop** as *sigma* increases.

Comparisons on Human3.6M



PoseFormerV2 suffers from **less performance drop** while being **more efficient**.

*the size of markers denotes computational cost

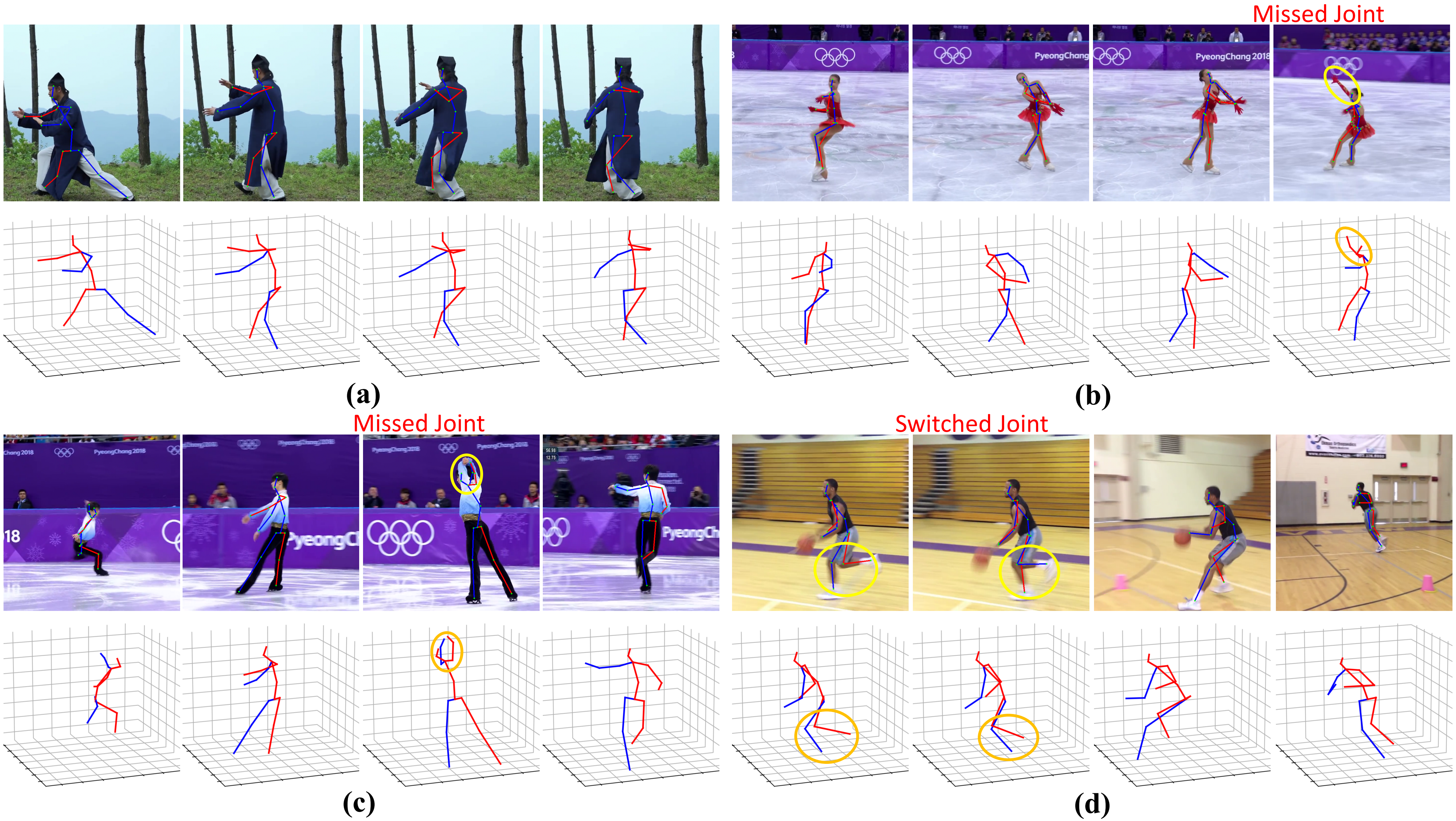
Comparisons on MPI-INF-3DHP

Method		PCK ↑	AUC ↑	MPJPE ↓
Mehta *et al.* [23]	3DV'17	75.7	39.3	117.6
Mehta *et al.* [24]	ACM ToG'17	76.6	40.4	124.7
Pavllo *et al.* [29] (T=81)	CVPR'19	86.0	51.9	84.0
Pavllo *et al.* [29] (T=243)	CVPR'19	85.5	51.5	84.8
Lin *et al.* [17] (T=25)	BMVC'19	83.6	51.4	79.8
Li *et al.* [14]	CVPR'20	81.2	46.1	99.7
Chen *et al.* [5] (T=81)	TCSVT'21	87.9	54.0	78.8
PoseFormerV1 [45] (T=9)(†)	ICCV'21	95.4	63.2	57.7
MHFormer [16] (T=9)	CVPR'22	93.8	63.3	58.0
MixSTE [43] (T=27)	CVPR'22	94.4	66.5	54.9
P-STMO [32] (T=81)(*)	ECCV'22	**97.9**	75.8	32.2
PoseFormerV2 (T=81)		**97.9**	**78.8**	**27.8**

PoseFormerV2 achieves the **state-of-the-art** performance on MPI-INF-3DHP

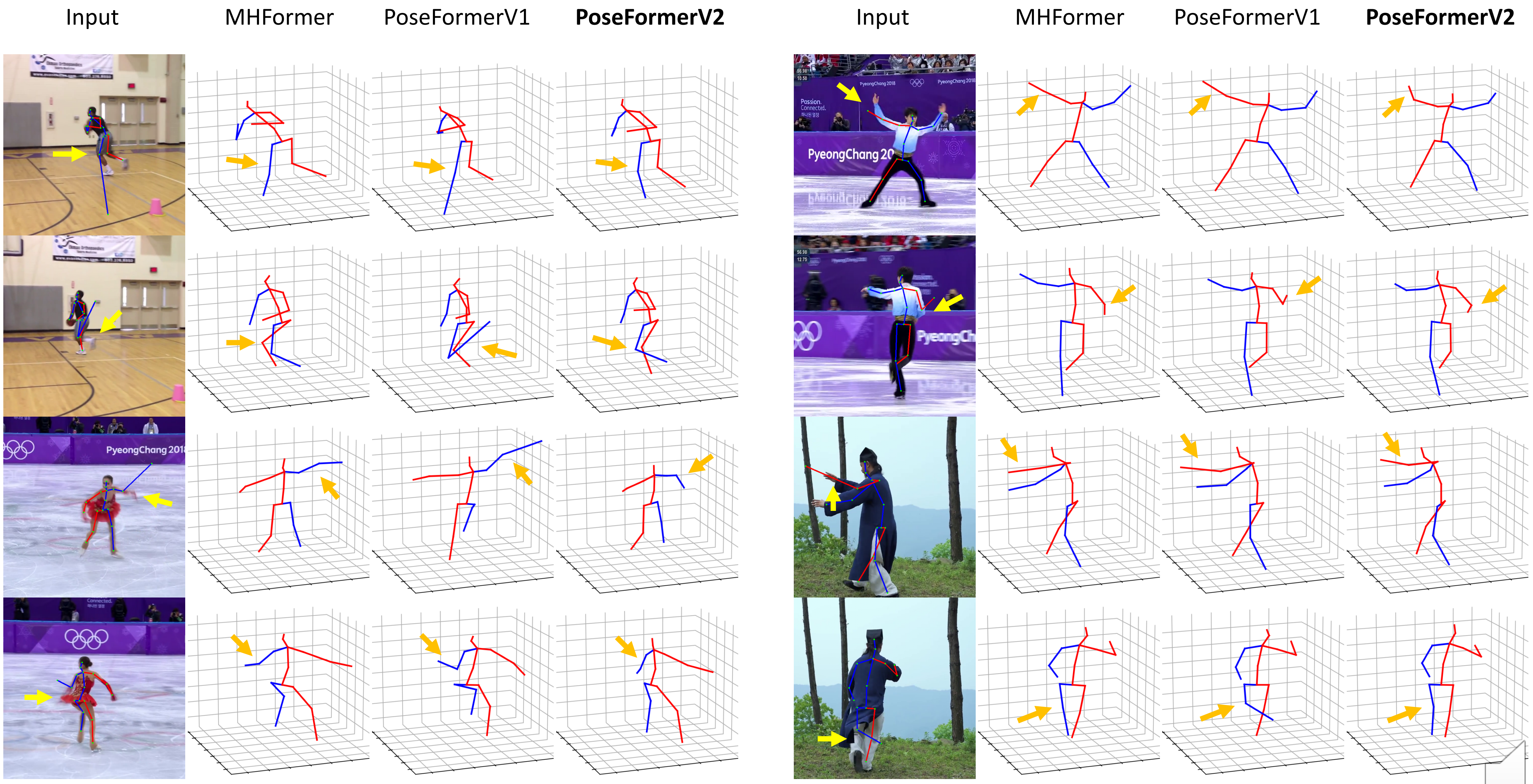
Qualitative Results

Challenging in-the-wild images



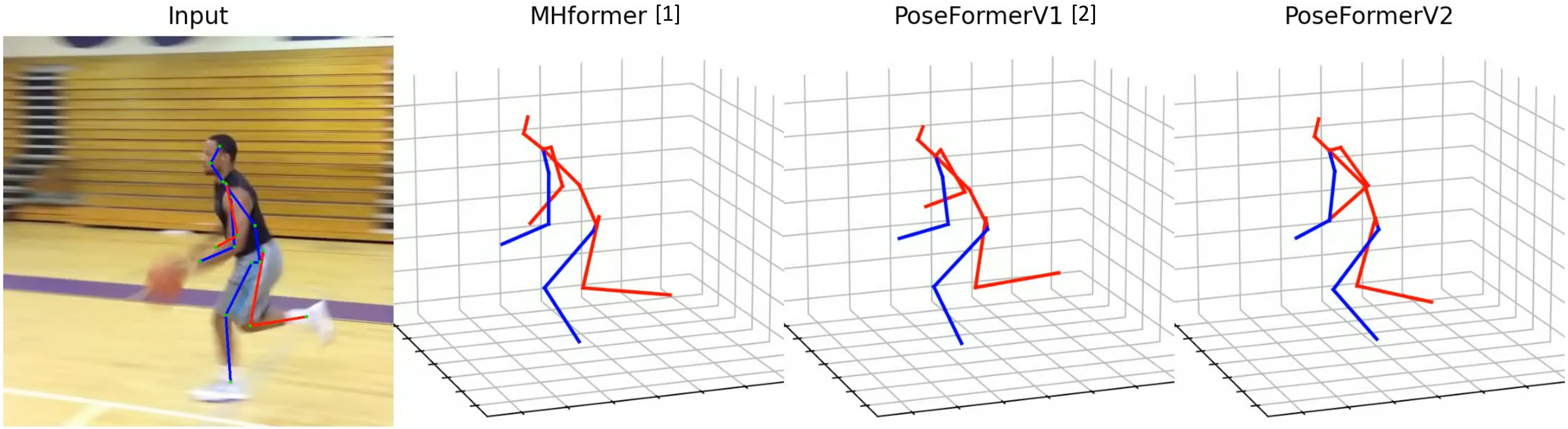
PoseFormerV2 infers correct 3D pose with unreliable 2D joint detection

We add Gaussian noise to a randomly-selected 2D joint to compare the robustness of models



PoseFormerV2 obtains reliable 3D pose with even highly deviated 2D joint detection

We add Gaussian noise to **all** 2D joints and **PoseFormerV2** shows a surprisingly good **temporal consistency**.



[1] Li et al. MHFormer: Multi-hypothesis transformer for 3d human pose estimation. In CVPR, 2022.

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

Thanks for Watching!

- **Project Page**
- (code & video):
- **qitaozhao.github.io/PoseFormerV2**

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA