Accurate and Steady Inertial Pose Estimation through Sequence Structure Learning and Modulation

Yinghao Wu[1] Chaoran Wang[1] Lu Yin[1] Shihui Guo[1] Yipeng Qin[2]

[1] School of Informatics, Xiamen University, China

[2] School of Computer Science & Informatics, Cardiff University, UK

XIAMEN UNIVERSITY

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

NEURAL INFORMATION PROCESSING SYSTEMS

Inertial Pose Estimation

IMU-Based: **environment-free** **occlusion-unaware** **privacy-friendly**

Native Transformer

Native Transformer

variable-length sequence

lacks inductive bias

+

Inertial Pose Estimation

fixed-length sequences

structural properties

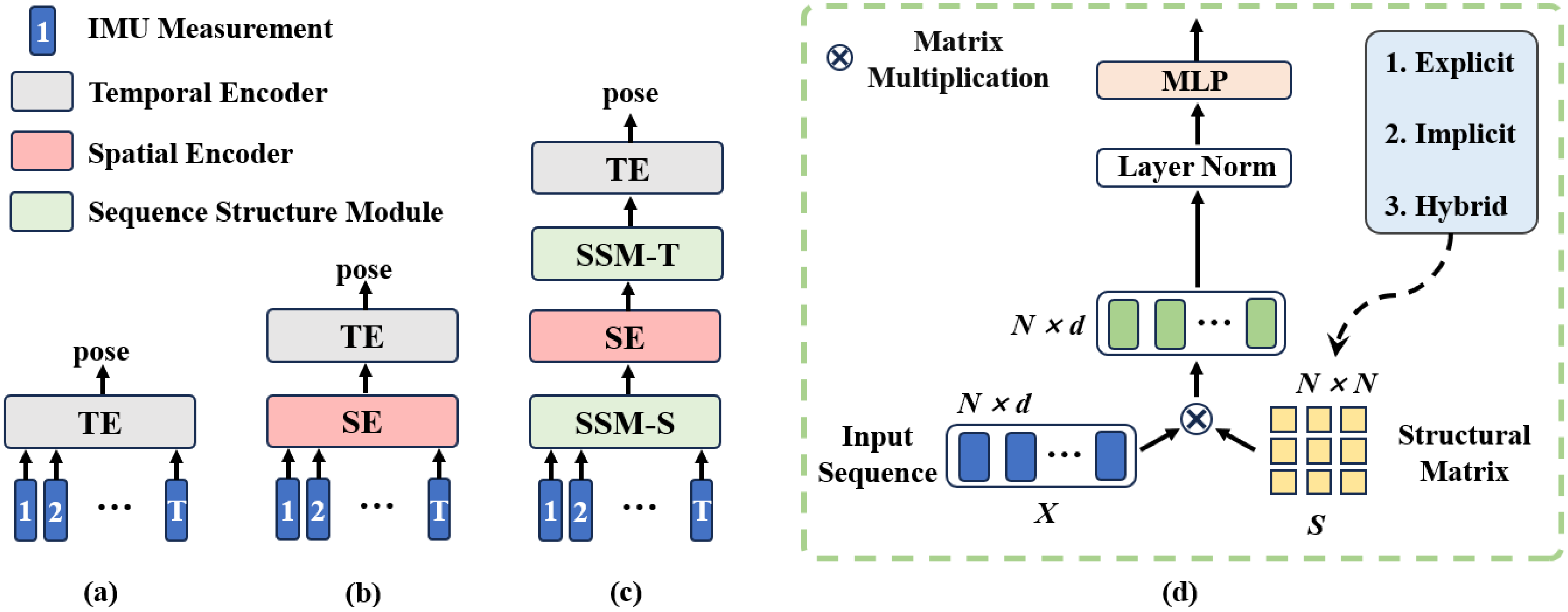
inaccurate postures

unacceptable jitter

Contribution1:

We identify a key limitation of the native transformer architecture: its lack of inductive biases for modeling **fixed-length sequences** with inherent structural properties. To address this shortcoming, we propose a novel Sequence Structure Module (SSM) that enables transformers to effectively capture and leverage the structural priors present in fixed-length sequential data.

Sequence Structure Module (SSM)



Contribution2:

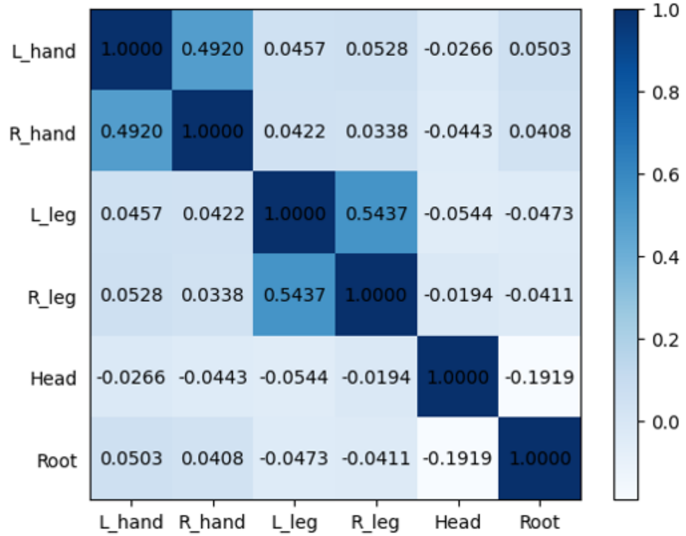
For inertial motion capture tasks involving sequential IMU data, we propose two SSM variants: SSM-S and SSM-T, which incorporate structural inductive biases of the IMU sensor layout (spatial) and time frames (temporal), respectively, into transformer learning.

Structural Matrix

Structural Matrix for SSM-S

$$C^k(i,j) = \frac{\mathrm{cov}(R_{(i)}^{(k,sub)}, R_{(j)}^{(k,sub)})}{\sqrt{\mathrm{var}(R_{(i)}^{(k,sub)}) \times \mathrm{var}(R_{(j)}^{(k,sub)})}}$$

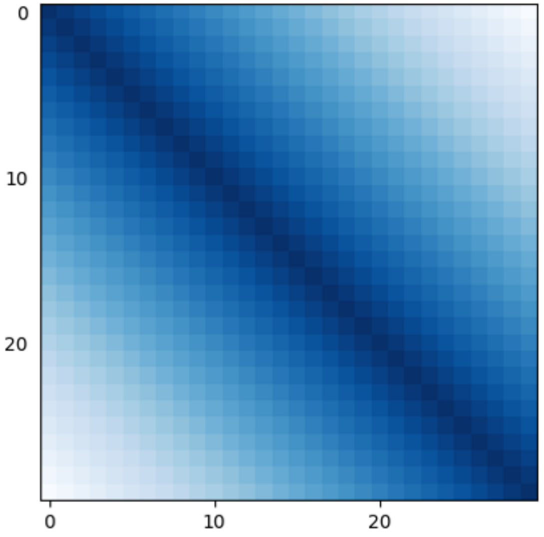
$$S_{E-S} = \frac{1}{3} \sum_{k=x,y,z} C^k$$



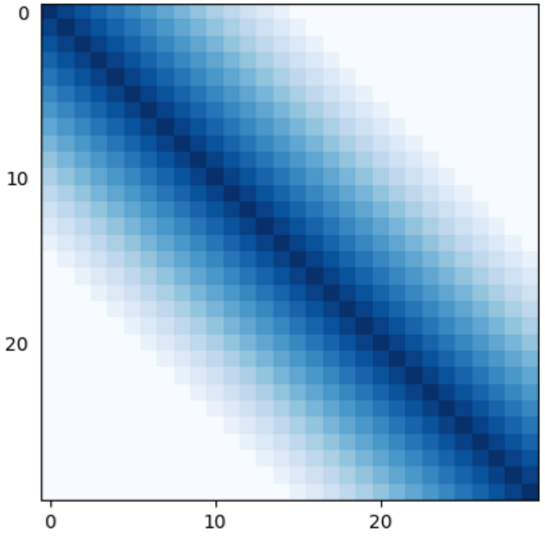
S_{E-S}

Structural Matrix for SSM-T

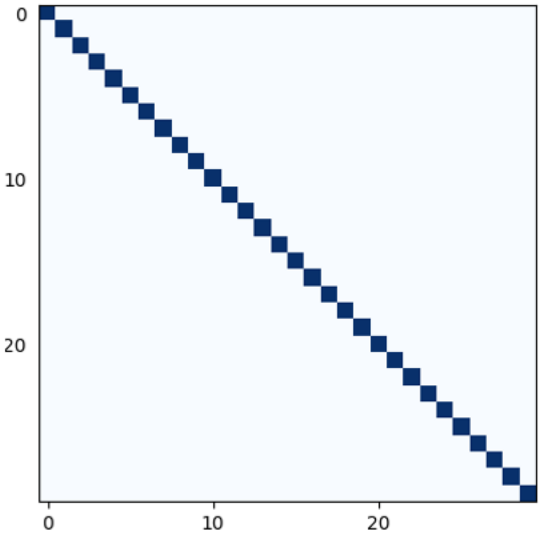
$$S_{E-T}(i,j) = \begin{cases} 0 & \text{if } |i-j| \geq \sigma, \\ 1 - \dfrac{|i-j|}{\sigma} & \text{else} \end{cases}$$



$S_{E-T}(\sigma = 30)$



$S_{E-T}(\sigma = 15)$



$S_{E-T}(\sigma = 1)$

Experimental Results

Table 1: Comparison with SOTA methods on DIP-IMU [18] and TotalCapture [46] datasets with SMPL [29] skeleton. **Bold** indicates best and <u>underline</u> indicates runner-up results.

	DIP-IMU					TotalCapture				
	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter
DIP[18]	17.10	15.16	7.33	8.96	3.01	18.62	17.22	9.42	11.22	3.62
Transpose[56]	17.03	8.86	6.03	7.14	1.08	16.40	12.77	6.42	7.20	1.83
TIP[20]	16.92	9.07	5.63	6.62	1.53	13.20	12.24	5.68	6.78	1.57
PIP[55]	15.02	8.72	5.01	6.02	<u>0.14</u>	12.93	12.04	5.61	6.51	<u>0.18</u>
DynaIP[60]	14.11	<u>7.00</u>	<u>4.97</u>	5.97	0.18	12.42	11.06	5.11	5.79	0.22
PNP[57]	<u>13.71</u>	8.75	<u>4.97</u>	<u>5.77</u>	0.17	<u>10.89</u>	<u>10.45</u>	<u>4.74</u>	<u>5.45</u>	0.26
Ours	**7.90**	**6.06**	**3.12**	**3.78**	**0.07**	**7.00**	**6.82**	**3.36**	**4.00**	**0.09**

Table 2: Comparison with SOTA methods on AnDy [33] and CIP [37] datasets with Xsens [41] skeleton.

	AnDy			CIP		
	SIP Err	Ang Err	Pos Err	SIP Err	Ang Err	Pos Err
Transpose[56]	12.15	6.29	4.91	20.06	8.75	6.86
TIP[20]	10.11	4.55	3.56	13.05	5.67	4.30
PIP[55]	9.49	4.09	<u>3.29</u>	12.68	5.52	4.12
DynaIP[60]	<u>8.93</u>	<u>3.45</u>	3.41	<u>11.42</u>	**4.54**	<u>3.69</u>
Ours	**4.56**	**3.37**	**1.73**	**8.14**	<u>5.49</u>	**2.57**

Table 3: Ablation study of SSM-S and SSM-T.

Models	Ang Err	Jitter	τ
Baseline	8.82	0.48	14.25
+ SSM-S	7.83	0.43	12.04
+ SSM-T	7.93	0.09	8.68
Ours	**6.82**	**0.09**	**7.46**

Experimental Results

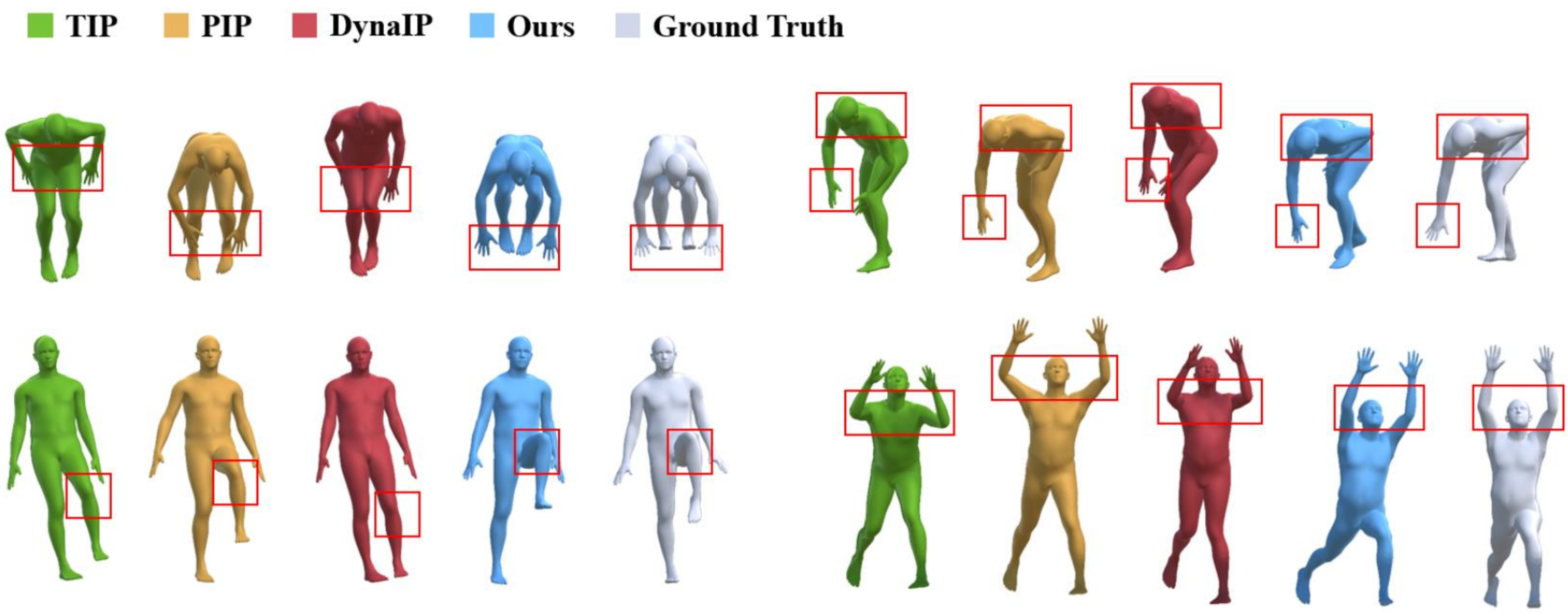


Figure 4: Qualitative comparisons with the state-of-the-art methods on TotalCapture dataset.

Experimental Results

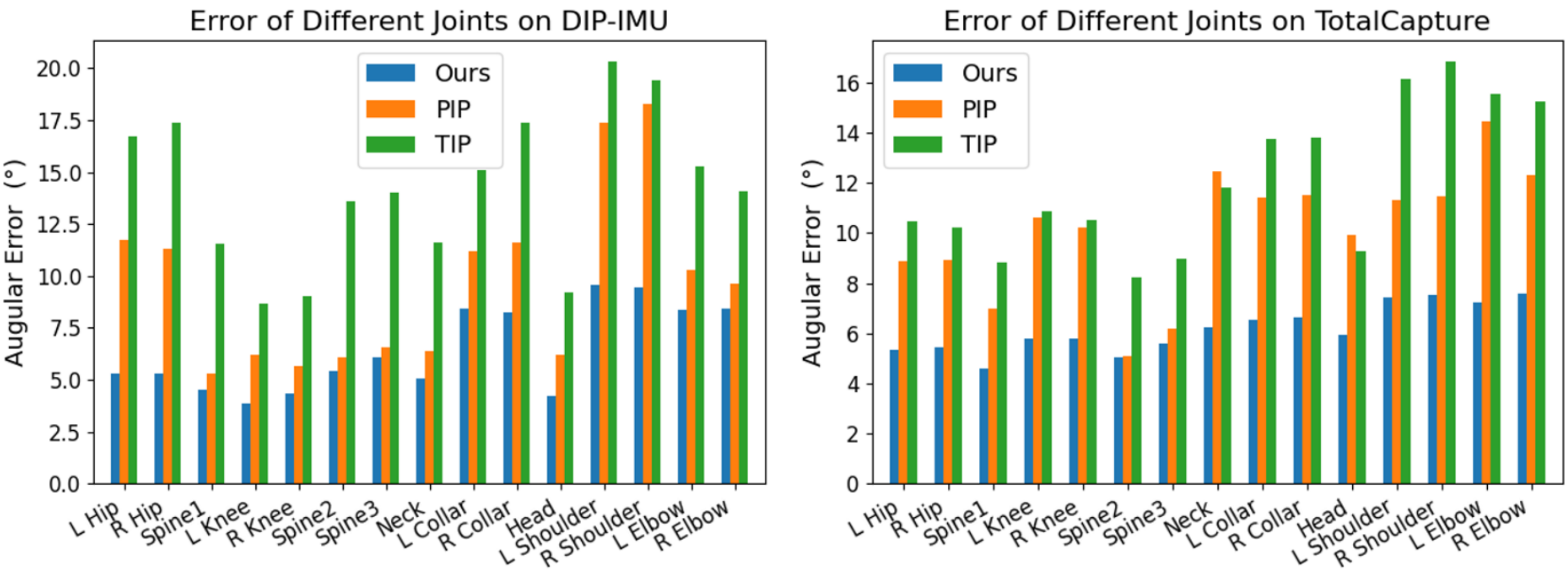


Figure 5: Error of different joints on DIP-IMU and TotalCapture datasets.

Live Demo

Accurate and Steady Inertial Pose Estimation through Sequence Structure Learning and Modulation

Thank you!