

Supplementary Material for APTpose: Anatomy-aware Pre-Training for 3D Human Pose Estimation

Qing-Wen Yang¹
ss109062702@gapp.nthu.edu.tw

Kai-Wen Duan¹
kevin77688@gapp.nthu.edu.tw

Ting-Yi Lu¹
grace1287986@gapp.nthu.edu.tw

Kevin Lin²
keli@microsoft.com

Cheng-Yen Yang³
cycyang@uw.edu

Lijuan Wang²
lijuanw@microsoft.com

Jenq-Neng Hwang³
hwang@uw.edu

Shang-Hong Lai¹
lai@cs.nthu.edu.tw

¹ Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

² Microsoft Corporation, USA

³ University of Washington, USA

In this supplemental material, we offer additional details to complement our proposed anatomy-aware pre-training model. First, the implementation details and network architecture are elaborated upon in Section A. Subsequently, we delve into the noising pipeline in Section B, elucidating the integration of image-based datasets into our sequence-based model. In Section C, we conduct an ablation study on various noise addition strategies to simulate sequence variations. Lastly, we present additional quantitative and qualitative evaluation results in Sections D and E, respectively.

A Detailed of implementation and Model Architecture

A.1 Implementation Details

We implement our proposed APTPose with PyTorch and conduct experiments using a computer equipped with two NVIDIA Tesla V100 GPUs. The standard data augmentations, such as horizontal flipping of poses, are applied to both 2D and 3D pose data. For model training, we utilize Adam optimizer [1] with a batch size of 160. The initial learning rates are set to $1e^{-4}$ and $9e^{-3}$ for pre-training and fine-tuning, respectively. The training process lasts for

80 epochs. Moreover, we also employ an exponential learning rate decay schedule with a decay factor of 0.97 to aid the training procedure.

Additionally, we set $\lambda_{3D}=1$ and $\lambda_{2D}=0.3$ to balance the effects of the 3D pose loss and the 2D reprojection loss during the pre-training stage. During the fine-tuning stage, the weighting factors are set to $\lambda_{mp}=1$, $\lambda_{mb}=0.5$, $\lambda_{sp}=1$ and $\lambda_{sb}=0.5$ to balance the effects of pose loss and bone vector loss for both multi-frame and single-frame scenarios.

A.2 Model Architecture

To validate the effectiveness of our training scheme, we conduct experiments using a model backbone similar to the previous pre-training approach P-STMO [14]. However, it is worth noting that our lifting model is flexible and can be replaced by any existing architecture commonly used in human pose estimation literature.

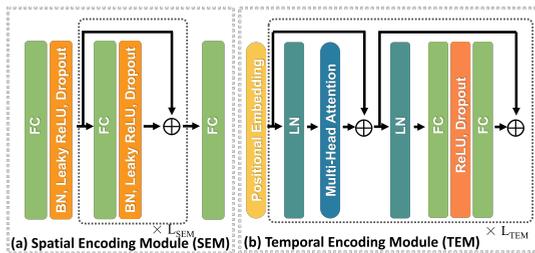


Figure A: **The detailed architecture of AaE, consisting of four SEMs and a TEM.** (a) Spatial Encoding Modules (SEM), (b) Temporal Encoding Module (TEM).

Anatomy-aware Encoder (AaE). The APTPose framework comprises an Anatomy-aware Encoder (AaE), that incorporates four Spatial Encoding Modules (SEMs), and a Temporal Encoding Module (TEM). Each SEM is designed to capture a specific body component of the human structure. For computational efficiency, SEM is constructed using a simple MLP block with residual connections, as illustrated in Figure A(a). TEM, shown in A(b), is constructed using a vanilla Transformer, consisting of multi-head self-attention and MLP block, allowing our model to capture long-range temporal dependencies. Note that in our approach, we employ an asymmetric encoder decoder design, where the depth of the decoder is lower than that of the encoder.

Reprojection Module (RM). The Reprojection Module (RM) is similar in structure to the SEM, using MLP blocks with residual connections as in Figure A(a). The difference is that the RM takes 3D poses as input and maps them to 2D poses.

Many-to-One Frame Aggregator (MOFA). Figure B illustrates the Many-to-One Frame Aggregator (MOFA), which is embedded before the final pose estimation at Stage II. MOFA follows a structure similar to the vanilla Transformer but replaces the MLP block with strided 1D convolution. This substitution facilitates the utilization of a full-to-single scheme, which aims to enhance temporal smoothness across the entire sequence scale and refine the representation at the single-frame level.

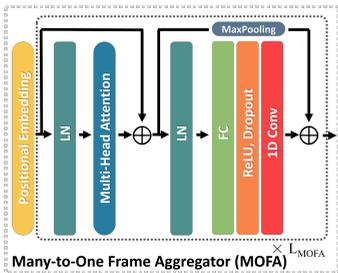


Figure B: The detailed architecture of MOFA.

COCO+N(0, 0.0005)	COCO+N(0, 0.005)	COCO+N(0, 0.05)
31.89	31.76	31.25

Table A: Ablation study on the effectiveness of adding different levels (std=0.5, 0.005 and 0.0005) of Gaussian noise to the annotations of COCO dataset to simulate sequence variations.

B Detailed of Noising pipeline

The limited availability and variability of public datasets with 2D-3D pose pairs present a common challenge, as these datasets are typically collected in controlled laboratory environments. To address this issue, utilizing existing in-the-wild 2D datasets is a viable option due to their closer approximation to real-world scenarios. Based on this rationale, we selected the well-known 2D image-based in-the-wild dataset, **COCO** [4]. Although COCO contains fewer data than H36M or 3DHP, its diverse range of human activities and variable image sizes make it a particularly challenging dataset for pose estimation models.

To bridge the gap between the image-based format and our sequence-based framework, we introduce a noising pipeline. This pipeline involves duplicating images into the required frames (e.g., 81 frames for 3DHP and 243 frames for H36M) and adding random Gaussian noise and jitter to pose annotations for sequence simulation. The extent of noise augmentation will be discussed in subsequent sections. Additionally, we leverage the 3D pseudo labels of the 2D dataset provided by [4], whose model also adopts an MLP as a backbone, similar to our reprojection module. This enables the creation of 2D-3D pairs of data to be fed into our model, thereby enhancing generalization through augmented training scenarios.

We conduct experiments on incorporating existing in-the-wild 2D dataset (e.g., COCO) and 3D dataset (e.g., H36M) for pre-training and seen it brings robust performance in in-the-wild 3DHP dataset. The quantitative result on corss dataset experiment also shown in D.2.

C Ablation Study for MixCOCO

Table A presents the results of our ablation experiments on the 3DHP dataset, aimed at determining the optimal amount of noise to include. In these experiments, we used 2D ground truth keypoints as input and evaluated performance under MPJPE.

MPJPE (GT)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
VideoPose [█] CVPR'19 (N=243)	35.2	40.2	32.7	35.7	38.2	45.2	40.6	36.1	48.8	47.3	37.8	39.7	38.7	27.8	29.5	37.8
Anatomy3D [█] TCSVT'21 (N=243)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.8
MixSTE [█] CVPR'22 (N=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
P-STMO [█] ECCV'22 (N=243)	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
GLA-GCN [█] ICCV'23 (N=243)	26.5	27.2	29.2	25.4	28.2	31.7	29.5	26.9	37.8	39.9	29.9	27.0	27.3	20.5	20.8	28.5
APTPose (N=243)	25.3	26.7	27.8	25.3	28.0	29.7	29.0	25.3	35.3	34.4	29.9	26.2	25.3	18.0	18.5	26.8
MPJPE (CPN)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
VideoPose [█] CVPR'19 (N=243)	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
Anatomy3D [█] TCSVT'21 (N=243)	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
MixSTE [█] CVPR'22 (N=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
P-STMO [█] ECCV'22 (N=243)	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
GLA-GCN [█] ICCV'23 (N=243)	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
APTPose (N=243)	38.4	41.3	39.7	39.7	44.9	50.6	40.3	40.3	56.1	60.0	43.7	41.8	42.3	29.7	29.6	42.6
APTPose+(Extra2D) (N=243)	38.4	41.2	40.4	40.0	45.2	50.7	40.0	40.8	55.2	60.3	43.7	40.9	42.9	29.2	29.4	42.5
P-MPJPE (CPN)	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
VideoPose [█] CVPR'19 (N=243)	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
Anatomy3D [█] TCSVT'21 (N=243)	33.1	35.3	33.4	35.9	36.1	41.7	32.8	33.3	42.6	49.4	37.0	32.7	36.5	25.5	27.9	35.6
MixSTE [█] CVPR'22 (N=243)	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
P-STMO [█] ECCV'22 (N=243)	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
GLA-GCN [█] ICCV'23 (N=243)	32.4	35.3	32.6	34.2	35.0	42.1	32.1	31.9	45.5	49.5	36.1	32.4	35.6	23.5	24.7	34.8
APTPose (N=243)	30.5	33.9	32.4	33.0	34.4	39.6	31.7	31.3	44.6	49.1	35.4	32.9	33.8	23.5	24.1	34.0
APTPose+(Extra2D) (N=243)	30.5	33.6	33.0	32.9	34.2	39.5	31.4	31.6	44.2	48.1	35.6	32.0	34.0	23.3	23.7	33.8

Table B: **Results on Human 3.6M** in millimeters (mm) under MPJPE and P-MPJPE (rigid alignment). **Top table:** result under MPJPE using 2D GT keypoints as input. **Middle & Bottom table:** results under MPJPE and P-MPJPE using CPN as 2D detector. The best results are highlighted in **Red**. The second-best are highlighted in **Blue**.

D Quantitative Result

D.1 Action-wise evaluation Result on H36M

Table B presents a detailed evaluation of APTPose’s performance, providing action-wise results compared to state-of-the-art methods on the Human3.6M dataset. In the upper part of Table B, we used 2D ground truth keypoints as input and evaluated performance in terms of MPJPE for the 2D-to-3D lifting task, assuming optimal 2D detector performance. The results show that APTPose surpasses the previous pre-training method [█] (26.8mm vs. 29.3mm) as well as other notable methods. To ensure a fair comparison, we also used the widely-used CPN as the 2D detector and evaluated APTPose under both MPJPE and P-MPJPE metrics, as shown in the middle and bottom sections of Table B, respectively. The results further confirm APTPose’s superior performance compared to [█] (42.6mm vs. 42.8mm in MPJPE, and 34.0mm vs. 34.4mm in P-MPJPE), as well as other promising methods.

As discussed in Section 4.2, while MixSTE demonstrates the lowest reconstruction error on the Human3.6M dataset, it incurs a significant computational overhead, exceeding other methods by over *200 times*. APTPose achieves competitive accuracy with significantly reduced computational demands (1.367 vs. 277.24 GFLOPs), as shown in Table 1a and Figure 3a. Additionally, APTPose surpasses GLA-GCN when using both ground truth and CPN 2D keypoints under the MPJPE metric, with lower complexity (1.367 vs. 1.558 GFLOPs).

Overall, APTPose demonstrates competitive performance across various scenarios while maintaining an appropriate level of complexity, underscoring its superior stability and generalization capabilities.

D.2 Cross-dataset experiment on 3DHP

To evaluate how APTPose facilitates model generalization to cross-scenario datasets, additional experiments are conducted as shown in Table C. The model is trained on the **H36M** dataset and tested on the **3DHP** dataset, where it is compared against various state-of-the-

Method	PCK↑	AUC↑	MPJPE↓	P-MPJPE↓
InterAug [10] ICASSP'22	81.6	48.2	93.4	-
PoseAug [9] CVPR'21	88.6	57.3	73.0	-
AdaptPose [8] CVPR'22	88.4	54.2	77.2	-
DynaBOA [4] T-PAMI'22	79.5	43.1	101.5	66.1
P-STMO [11] ECCV'22	86.9	51.9	86.9	58.1
APTPose	89.0	57.5	76.8	57.7

Table C: Cross-dataset evaluation on **3DHP** dataset. The best results are highlighted in **Red**. The second-best are highlighted in **Blue**.

art methods. Extensive experiments show that APTPose achieves significant improvements over [11] in PCK, AUC, MPJPE, and P-MPJPE metrics by 2.4%, 10.7%, 11.6%, and 0.7%, respectively. The result highlights the effectiveness of APTPose’s hierarchical masking strategy, emphasizing the learning of the human skeletal structure, for application across diverse scenarios compared to previous approaches that focus on learning individual joint nodes.

Notably, our approach even outperforms data augmentation-based methods [9, 10], as well as adaptation approaches [8, 4], achieving state-of-the-art performance on PCK, AUC, P-MPJPE metrics without utilizing data from the test set of the target domain.

E Qualitative Result

As shown in Figures C-G, we present more qualitative comparisons between our proposed APTPose and P-STMO [11] on five challenging in-the-wild videos, including *skating*, *basketball*, *dancing*, *rollerskate*, and *ski*. The results demonstrate that our approach significantly outperforms P-STMO in capturing realistic actions in real-world scenarios and predicting human proportions more reasonably.

References

- [1] Tianlang Chen, Chen Fang, Xiaohui Shen, Yiheng Zhu, Zhili Chen, and Jiebo Luo. Anatomy-aware 3d human pose estimation with bone-based pose decomposition. *arXiv preprint arXiv:2002.10322*, 2020.
- [2] Ziyi Chen, Akihiro Sugimoto, and Shang-Hong Lai. Learning monocular 3d human pose estimation with skeletal interpolation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4218–4222. IEEE, 2022.
- [3] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020.
- [4] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang. Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2022.

- [5] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021.
- [6] Shanyan Guan, Jingwei Xu, Michelle Zhang He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5070–5086, 2022.
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [9] Dario Pavullo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019.
- [10] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022.
- [11] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8818–8829, 2023.
- [12] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13232–13242, 2022.

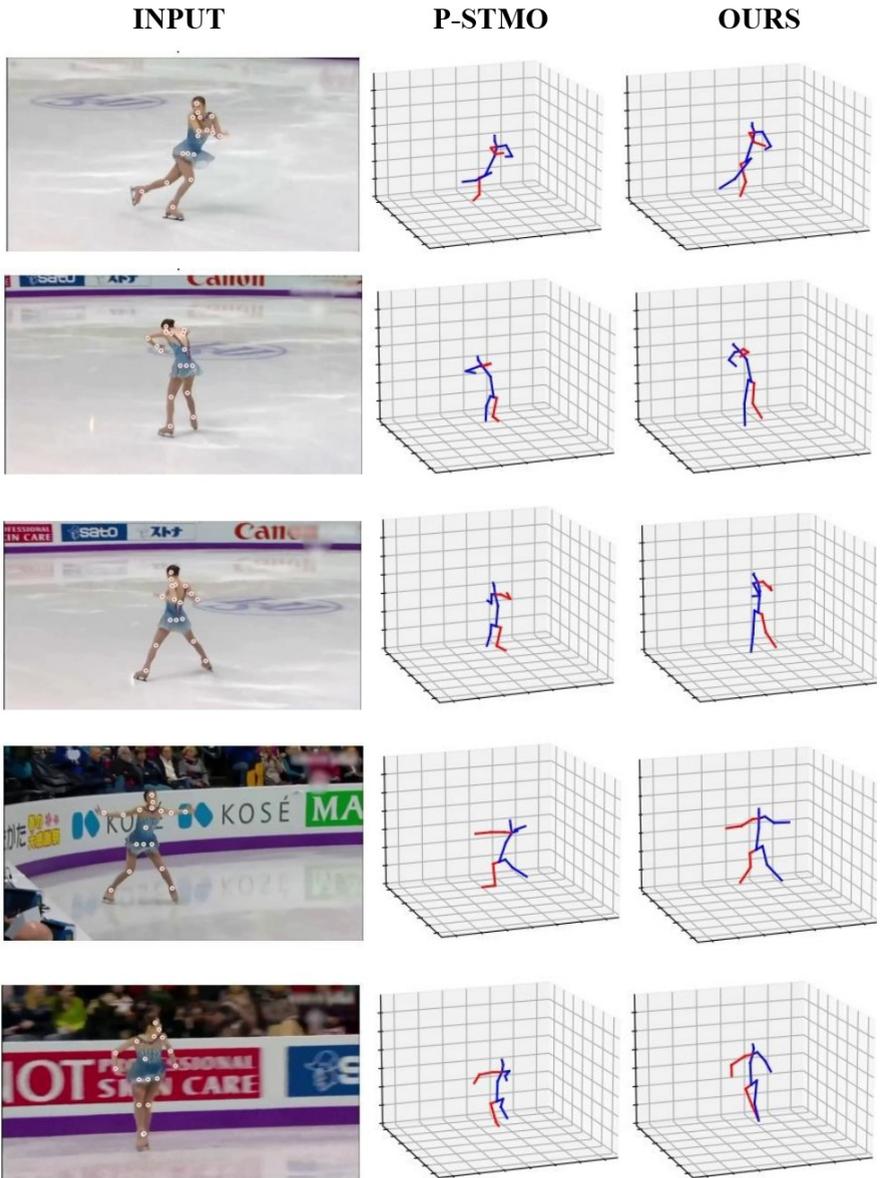


Figure C: Qualitative results on in-the-wild videos - **skating videos**. We compared our approach with the state-of-the-art method P-STMO [10].

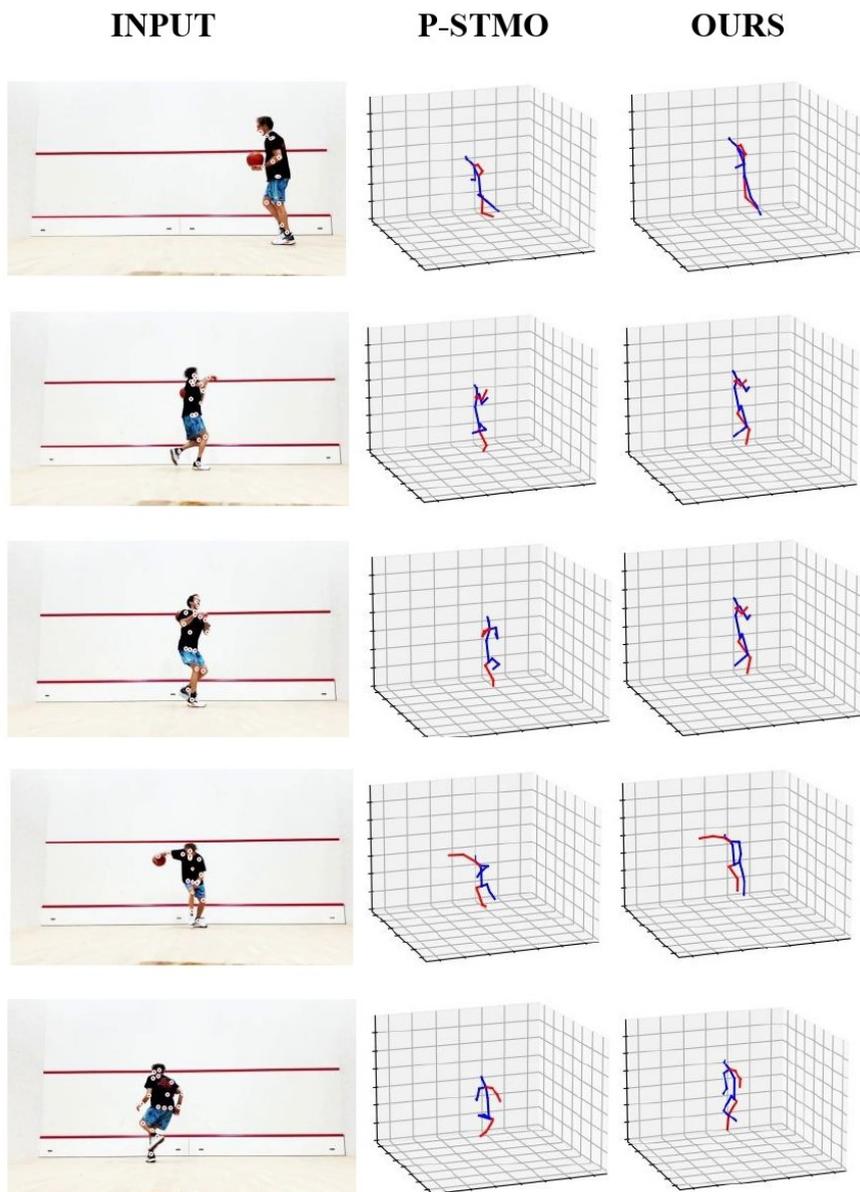


Figure D: Qualitative results on in-the-wild videos - **basketball videos**. We compared our approach with the state-of-the-art method P-STMO [14].

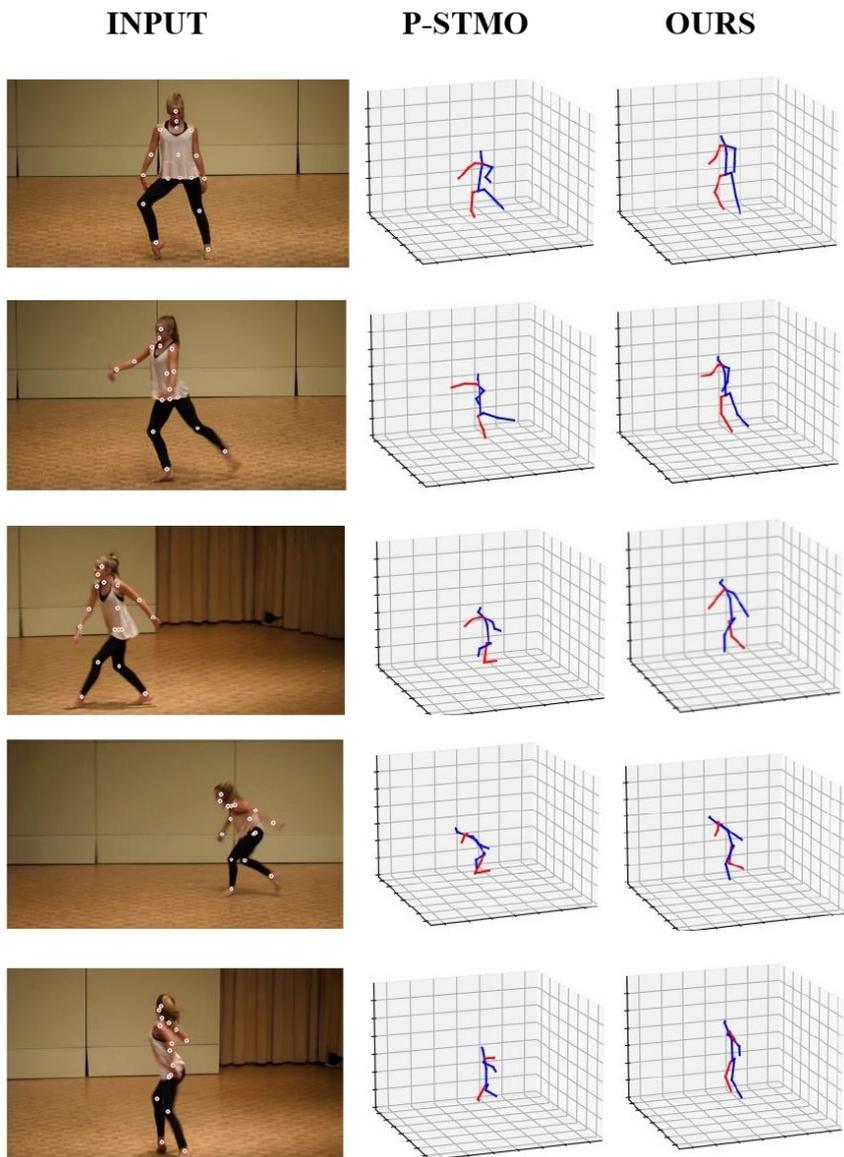


Figure E: Qualitative results on in-the-wild videos - **dancing videos**. We compared our approach with the state-of-the-art method P-STMO [10].

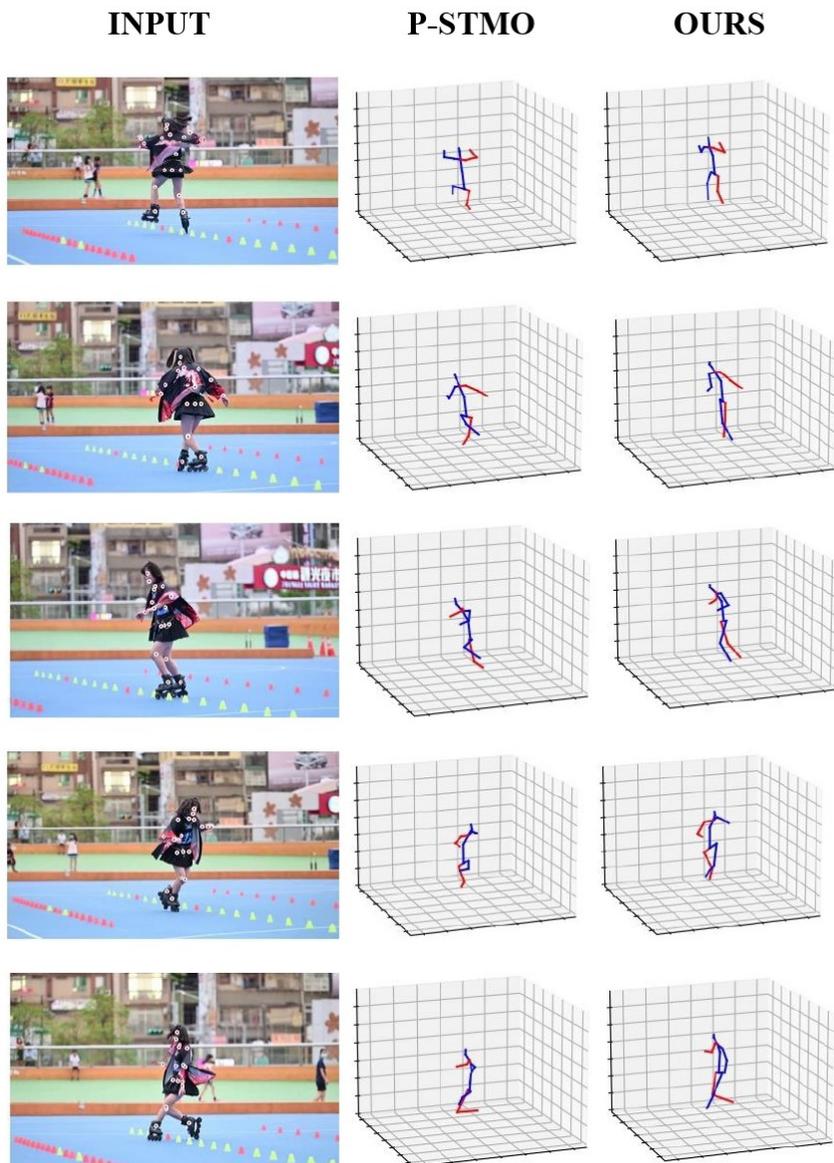


Figure F: Qualitative results on in-the-wild videos - **rollerskate videos**. We compared our approach with the state-of-the-art method P-STMO [11].

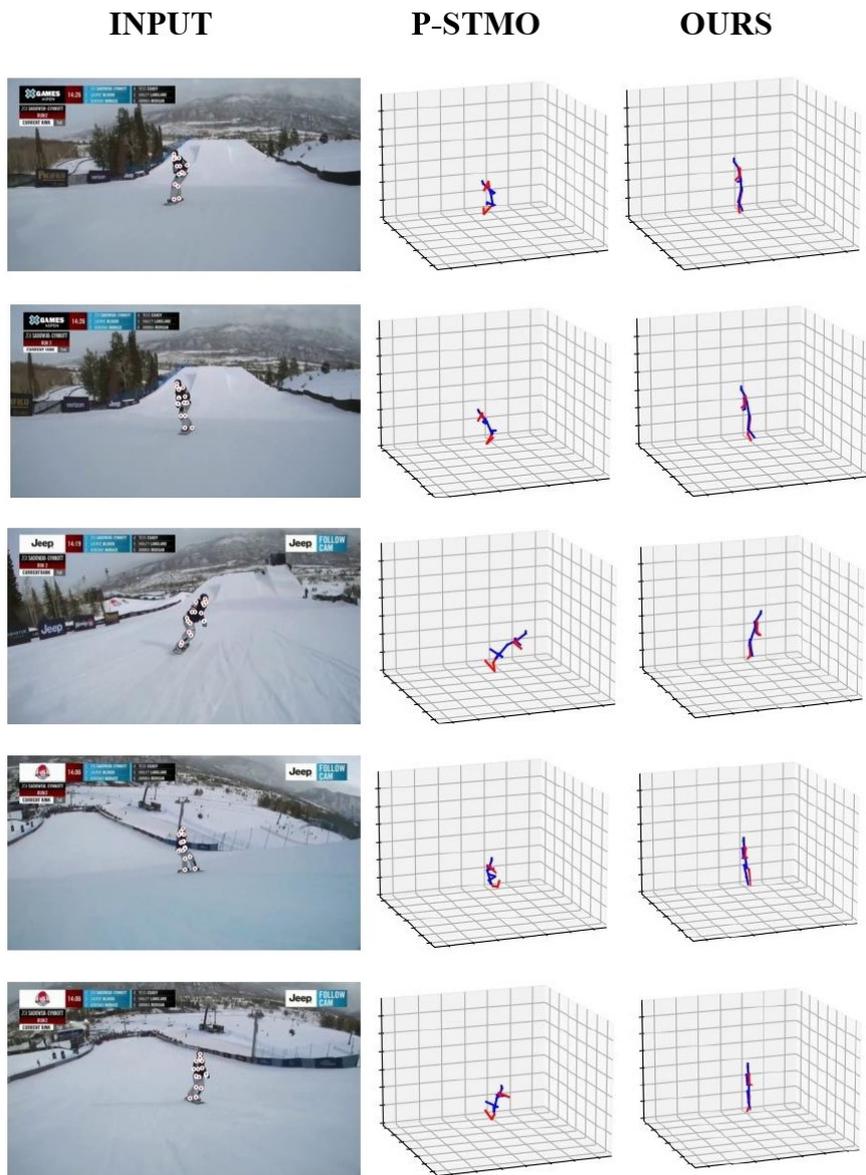


Figure G: Qualitative results on in-the-wild videos - **ski videos**. We compared our approach with the state-of-the-art method P-STMO [10].