PS-Mamba: Spatial-Temporal Graph Mamba for Pose Sequence Refinement

Supplementary Material

Contents

We provide comprehensive supplementary materials to support the details of the proposed PS-Mamba. **Section** 1 covers the architecture details, while **Section** 2 describes the training schemes, inference processes, baselines, datasets, and metrics. **Section** 3 presents a comparison between our approach and additional recent methods, **Section** 4 presents qualitative results to further verify the performance, and **Section** 5 provides the details for the ablation study. Lastly, **Section** 6 discusses limitations and future work.

We have uploaded the source code for reproducibility and welcome your feedback.

1. Architecture Details

The proposed PS-Mamba is the first framework that refines human pose sequences by integrating spatial-temporal graph learning with state space modeling. The core component is the ST-GSS block, which combines the ST-Graph and ST-SSM, using our effective GST-Scanning.

ST-GSS Block. The Spatial-Temporal Graph State Space (ST-GSS) block is the core component of PS-Mamba, designed to capture spatial-temporal relationships in human pose sequences. It integrates two key elements: the ST-Graph network and the ST-SSM module. This block processes human pose sequences by leveraging spatial relationships between joints across time and refining motion coherence using state space modeling. The block contains a residual skip connection that enhances the overall performance by preserving and improving both spatial and temporal features throughout the sequence.

GST-Scanning. We propose GST-Scanning (Graph-guided Spatial-Temporal Scanning), specifically designed for refining human pose sequences. It utilizes four scanning sequences to fully capture interactions between joints across both spatial and temporal dimensions. Specifically, the four scanning sequences are divided into two categories, each with two forward and two backward scans, all guided by the human skeleton graph.

- **Bidirectional Spatial Scanning**: This scans the joints in the spatial space, capturing the relationships among neighboring joints in the same frame.. Using a bidirectional approach captures both forward and backward dependencies among joints, enabling model to capture full spatial structure of the pose
- **Bidirectional Temporal Scanning**: This scans the joints across consecutive frames in the temporal domain. The bidirectional scanning enables the model to capture both past and future temporal dependencies between joints, ensuring the model effectively captures the evolving dynamics of joint motion over time.

By combining these four scanning sequences, GST-Scanning effectively captures intricate spatial-temporal graph relationships, improving the model's capacity to refine human pose sequences while ensuring enhanced coherence across both spatial and temporal dimensions.

ST-Graph. The ST-Graph integrates both Graph Convolutional Network (GCN) and Temporal Convolutional Network (TCN) layers to model spatial and temporal dependencies of human poses. It captures the spatial relationships between joints across consecutive frames by modeling the human pose as a graph where joints are nodes and spatial-temporal edges encode their relationships. Spatial edges are determined by the pose structure, while temporal edges link joints across frames. A dynamic graph weight matrix adaptively adjusts the influence of neighboring joints, addressing challenges such as jitter and ambiguity in human pose sequence refinement.

ST-SSM. The Spatial-Temporal State Space Model (ST-SSM) refines the features from the ST-Graph network through SSM blocks, applying layer normalization, depthwise convolutions, and selective 2D scanning (SS2D [13]) to capture spatial and temporal dependencies. This process smooths the temporal evolution of joint positions while preserving the spatial structure, enhancing both accuracy and motion coherence in the pose sequence.

Encoder and Decoder. We employ one FC layer for both encoding and decoding. The encoder and decoder are essential for processing and refining human pose sequences.

The encoder extracts high-dimensional features from the input sequence using pre-trained models. These features are then processed through multiple ST-GSS blocks to capture spatial-temporal dependencies and refine motion coherence. The decoder generates the refined human pose sequences from the processed features. This architecture ensures smooth pose refinement, addressing challenges such as jitter and ambiguity in human pose refinement.

2. Training and Inference Details

2.1. Training Schemes

The proposed PS-Mamba was trained on an NVIDIA RTX A6000 with a batch size of 128 for approximately 30 epochs. For the Human3.6M dataset and Denoising tasks, a learning rate of 1×10^{-3} was used, while for AIST++ and 3DPW datasets, the learning rate was set to 1×10^{-2}, with a learning rate decay factor of 1.0. The sliding window size was set to 32. The loss weights for MPJPE, PA-MPJPE, and Accel were 1.0, 4.0, and 0.1, respectively. Additionally, for the SMPL representation, we incorporated extra SMPL parameter losses, including a 6D pose loss, with a weight of 1.0. The Adam optimizer was employed with the AMS-Grad variant set to True. For the denoising task training, we adopt the noise addition strategy from SynSP [20] and GFPose [3], applying both Gaussian and uniform noise as used in the baseline methods. Using the same noise configuration as baseline methods ensures consistent training and a fair comparison with existing denoising approaches. For 2D representation on the Human3.6M [6] dataset, the input 3D is initialized from Hourglass [15]. For 3D representation on the Human3.6M [6] dataset, the input 3D is initialized from FCN [14]. For SMPL representation on the AIST++ [12, 17] dataset, the input SMPL parameters are initialized from SPIN [11].

2.2. Inferences Schemes

For a fair comparison, we adopt the inference setting from SmoothNet [23]. This ensures that the evaluation of PS-Mamba is consistent with the established protocols and comparable to existing methods in the field. Specifically, we use the same input processing, post-processing steps, and evaluation metrics as in SmoothNet to evaluate our method's performance For the denoising task, we follow the evaluation protocol from SynSP [20] and GFPose [3]. This includes the same dataset splits, input noise levels, and quantitative metrics used in these methods to facilitate an unbiased comparison of performance. The consistency in the evaluation procedures allows us to reliably measure improvements and demonstrate the effectiveness of PS-Mamba in various pose refinement tasks.

2.3. Baselines

SmoothNet [23] targeted temporal-only refinement, reducing jitter and enhancing accuracy.

SynSP [20] balanced smoothness and precision in both single- and multi-view settings by aligning similar poses across views for consistency.

MoManifold [4] used a neural distance field and a joint acceleration manifold to model realistic motion dynamics.

One-Euro [2] reduces jitter in real time using an adaptive cutoff frequency that adjusts to signal dynamics.

Gaussian1d [21] convolves the signal with a Gaussian function, minimizing group delay.

Savitzky-Golay [16] applies a polynomial fit within a local window, preserving features while reducing noise.

2.4. Datasets

Human3.6M [6] is a large-scale dataset for 3D human pose estimation, containing approximately 3.6 million frames from 11 subjects performing 15 actions, captured from 4 camera views. Recorded at 50 fps, it provides accurate 3D joint annotations from a motion capture system and camera parameters. Five subjects (S1, S5, S6, S7, S8) are used for training, and 2 (S9, S11) for testing.

AIST++ [12, 17] is a lage multi-modal dataset based on the AIST Dance Video DB, including 1,408 3D dance sequences (captured at 60 fps), along with 3D keypoint labels and calibrated camera parameters spanning over 10 million frames. It contains 30 subjects, 9 views, and 10 dance genres with synchronized music, split into about 5.86M training and about 2.85M testing frames.

3DPW [19] is a widely used 3D human pose dataset, including 57,682 frames of accurate 3D human poses and SMPL parameters at 30 fps, with 22,463 training frames and 35,219 testing frames. It contains challenging outdoor scenes, such as walking and climbing stairs, and is used to evaluate image-based or video-based methods under occlusion and complex real-world scenarios.

CMU-Mocap [18] includes a wide range of common activities, such as walking, swimming, and climbing, excluding flying and rock climbing. A total of 2,241,016 frames are used for training, while 755,857 are allocated for testing.

2.5. Metrics

We adopt the evaluation setup from SmoothNet [23] and measure the performance of PS-Mamba using three standard metrics: MPJPE [6], PA-MPJPE [1, 8], and Accel [9].

MPJPE (Mean Per Joint Position Error) [6] measures the average Euclidean distance between predicted and ground truth joint positions, expressed in millimeters (mm). It is commonly used to evaluate 3D pose estimation accuracy.

PA-MPJPE (Procrustes Aligned Mean Per Joint Position Error) [1, 8] measures the average Euclidean distance between predicted and ground-truth joint positions after performing a rigid alignment (scale, rotation, and translation), using Procrustes Analysis (PA) [5]. It calculates the Euclidean distance between aligned positions and is expressed in millimeters (mm), widely used in 3D human pose estimation evaluation.

Accel (Acceleration error) [9] is the average difference between estimated human pose along with ground-truth per-joint acceleration, reported in mm/s².

3. Comparison with more SOTAs

To conduct a more comprehensive evaluation of PS-Mamba against the recent methods, as shown in Table 1, we present an additional comparison with DeciWatch [22] and HANet [7] on the Human3.6M and 3DPW datasets. Both DeciWatch [22] and HANet [7] leverage a sliding window of size 100, which contributes to their relatively good acceleration error performance. However, PS-Mamba consistently achieves the best results, as evidenced by its lower MPJPE and acceleration error. This superiority is attributed to the unique combination of spatial-temporal graph modeling and state-space modeling, which enables PS-Mamba to effectively handle and refine human pose sequences, even with the smaller sliding window size of 32. The ability of PS-Mamba to maintain high accuracy and smoothness in complex scenarios highlights the strength of its advanced modeling techniques.

Dataset	Method	MPJPE↓	Accel↓
Human3.6M	FCN [14]	54.6	19.2
	FCN [14] + DeciWatch [22]	52.8	1.5
	FCN [14] + HANet [7]	51.8	2.0
	FCN [14] + SmoothNet [23]	52.7	1.0
	FCN [14] + SynSP [20]	51.4	1.0
	FCN [14] + PS-Mamba (Ours)	**49.6**	**0.9**
3DPW	PARE [10]	79.0	25.6
	PARE [10] + DeciWatch [22]	77.2	6.9
	PARE [10] + HANet [7]	77.1	6.8
	PARE [10] + SmoothNet [23]	78.1	5.9
	PARE [10] + SynSP [20]	76.2	6.2
	PARE [10] + PS-Mamba (Ours)	**75.4**	**5.8**

Table 1. Comparison of our PS-Mamba with additional state-of-the-art methods on the Human3.6M [6] and 3DPW [19] datasets. Note that DeciWatch [22] and HANet [7] use a large window size of 100, while our method utilizes a sliding window size of 32.

4. Qualitative Comparison and Analysis

Due to the page limitations of the main manuscript, we present qualitative visual results in this section to explicitly demonstrate PS-Mamba's improvements in refining human pose sequences, especially challenging scenarios.

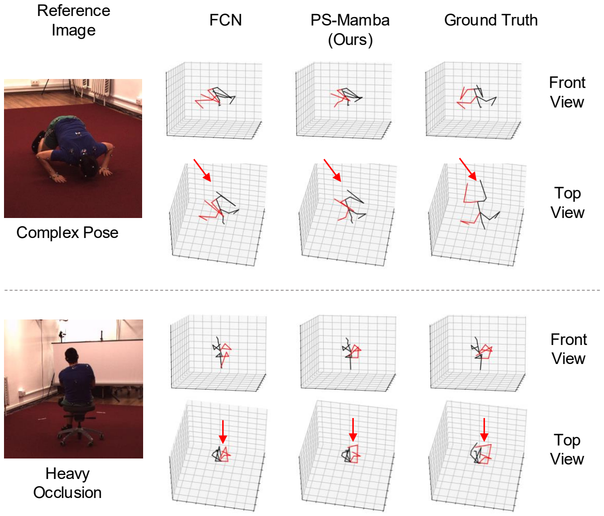

Figure 1. **Failure cases**. When faced with complex poses or heavy occlusions, both our method and the compared method fail to achieve the accurate human pose.

Visual Comparison. Figure 2 illustrates a comparison of our method with FCN [14] on the Human3.6M [6] dataset. As shown in Figure 5, Figure 6, and Figure 7, our PS-Mamba outperforms SynSP [20] and SmoothNet [23] in 3D mesh accuracy on AIST++ [12, 17]. PS-Mamba demonstrates superior accuracy in resolving ambiguous poses. For example, in the first row, FCN [14] misidentifies the left and right feet, while PS-Mamba generates poses that are closely aligned with the Ground Truth. Similarly, in the fifth row, our model demonstrates robust performance even in occlusions. These results verify PS-Mamba's effectiveness in handling ambiguity and maintaining robustness in challenging scenarios.

Acceleration Error and MPJPE Analysis. Figures 3 and 4 further compare PS-Mamba with VIBE [9] on the AIST++ [12, 17] dataset and FCN on the Human3.6M dataset. PS-Mamba achieves consistently lower acceleration error and MPJPE, demonstrating its capability to produce smoother motion trajectories and more accurate poses. These visualizations emphasize its ability to enhance motion smoothness while maintaining high precision, especially under noisy or complex input pose sequences.

Jitter Analysis. To analyze the impact of jitters and better highlight the improvements, we provide comparison videos visualizing the refined 3D pose results on the AIST++ dataset. These videos display the pose skeleton sequences along with corresponding acceleration errors and MPJPE. For single-frame methods, we use FCN [14], while video-based methods employ VIBE [9]. The visualizations reveal that jitters are often unbalanced, with most frames exhibiting minor jitters, while long-term jitters lead to substantial errors. PS-Mamba reduces both short- and long-term motion jitters, leading to notable improvements in smoothness.

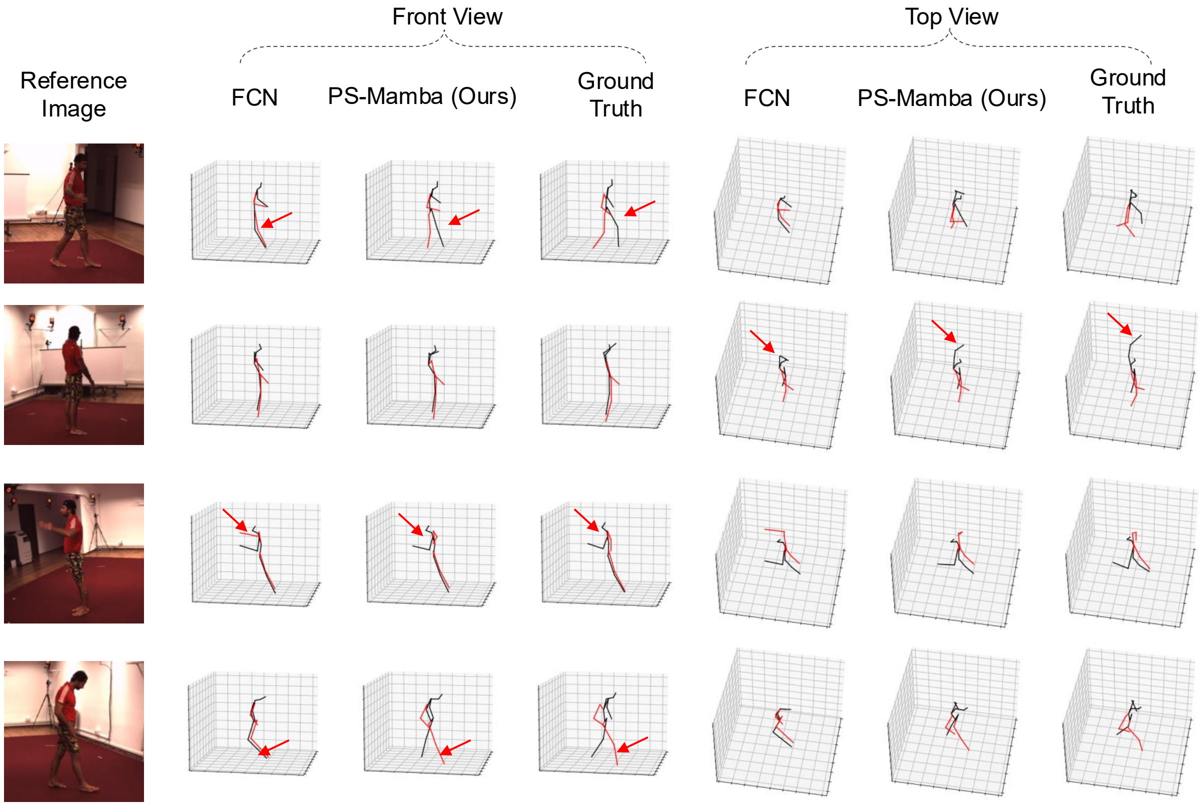

Figure 2. Qualitative comparison on the Human3.6M [6] Dataset.

These comparisons and analyses highlight PS-Mamba's effectiveness in pose refinement, demonstrating its robustness and accuracy across various datasets and scenarios.

5. Ablation Details

w/o ST-GSS: This excludes the Spatio-Temporal Graph-guided State Space (ST-GSS) block, which integrates spatial and temporal information for effective pose refinement.
w/o ST-SSM: This removes the Spatio-Temporal State-Space Module (ST-SSM), which is responsible for modeling temporal dynamics in pose sequences.
w/o ST-Graph: This ablation eliminates the Spatio-Temporal Graph, which models both spatial relationships among joints, enhancing the understanding of dynamic human pose sequence.
w/o Weight $\boldsymbol{W}$: This configuration excludes the graph weight matrix $\boldsymbol{W}$, which assigns learned importance to connections between nodes in the spatio-temporal graph.
w/o Temporal: This omits the temporal strategy, which models dependencies for motion temporal consistency.
w/o Residual: This removes the residual connections, which bypasses layers and adds input directly to the output.
w/o MPJPE Loss: This ablation excludes the Mean Per Joint Position Error (MPJPE) loss, a key supervision loss for accurate 3D pose estimation.
w/o PA-MPJPE Loss: This removes the Procrustes-Aligned MPJPE (PA-MPJPE) loss, calculated from pose alignment after Procrustes-Aligned transformation.
w/o Accel Loss: This configuration omits the acceleration loss, designed to ensure smooth temporal transitions and minimize jitter in motion dynamics.

6. Limitations and Future Work

While PS-Mamba demonstrates significant improvements in pose refinement, it has limitations in handling highly complex or out-of-distribution motion sequences. As shown in Figure 1, our method faces challenges in handling complex poses and heavy occlusion, which can impact the accuracy of pose refinement in such scenarios. The model's performance is heavily dependent on the specific training data, which may limit its ability to generalize to diverse or extreme motion patterns not represented in the dataset. To address this, future work could focus on training a more generalized model using larger and more diverse motion datasets, enabling PS-Mamba to adapt to a wider range of human poses and dynamic movements. This would enhance the model's robustness and applicability in real-world scenarios with varied human motion types.

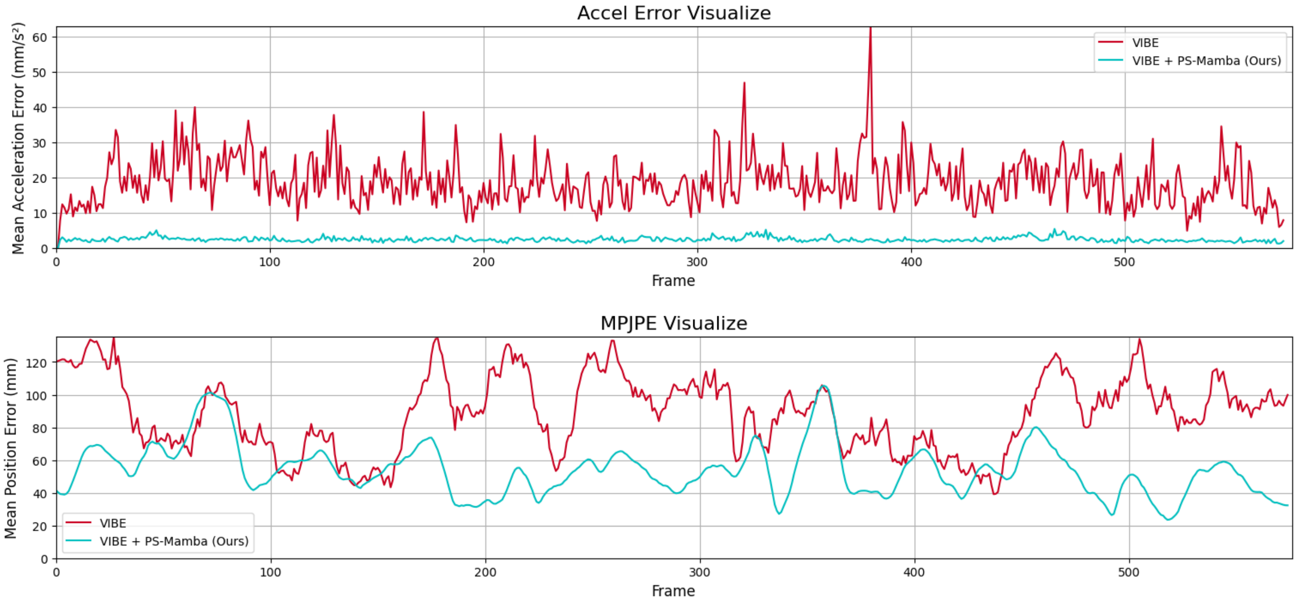

Figure 3. **Comparison of Acceleration Error and MPJPE on the AIST++ [12, 17] Dataset.** This figure displays sampled visualizations of acceleration error and MPJPE across multiple frames on the AIST++ [12, 17]. Our PS-Mamba method consistently outperforms VIBE [9], achieving the lowest acceleration error and MPJPE. These results highlight PS-Mamba's ability to generate smoother motion trajectories and more accurate poses, demonstrating its robustness in enhancing motion smoothness and maintaining precision under challenging conditions. This demonstrates PS-Mamba's robustness as an effective solution for human pose sequence refinement.

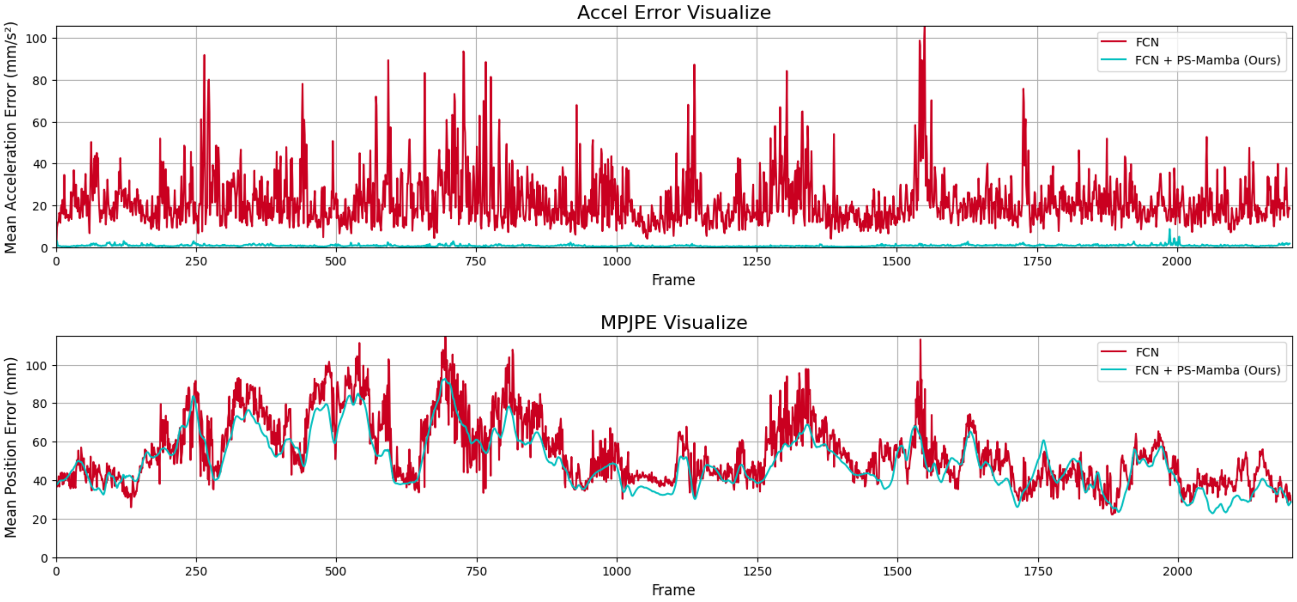

Figure 4. **Comparison of Acceleration Error and MPJPE on the Human3.6M [6] Dataset.** This figure presents sampled visualizations of acceleration error and MPJPE across multiple frames on the Human3.6M [6]. Our PS-Mamba method consistently outperforms FCN [14], achieving the lowest acceleration error and MPJPE. This demonstrates PS-Mamba's capability to produce smoother motion trajectories and more precise poses. The results highlight its robustness in addressing motion smoothness and maintaining accuracy under challenging scenarios, establishing it as a reliable solution for human pose sequence refinement task.

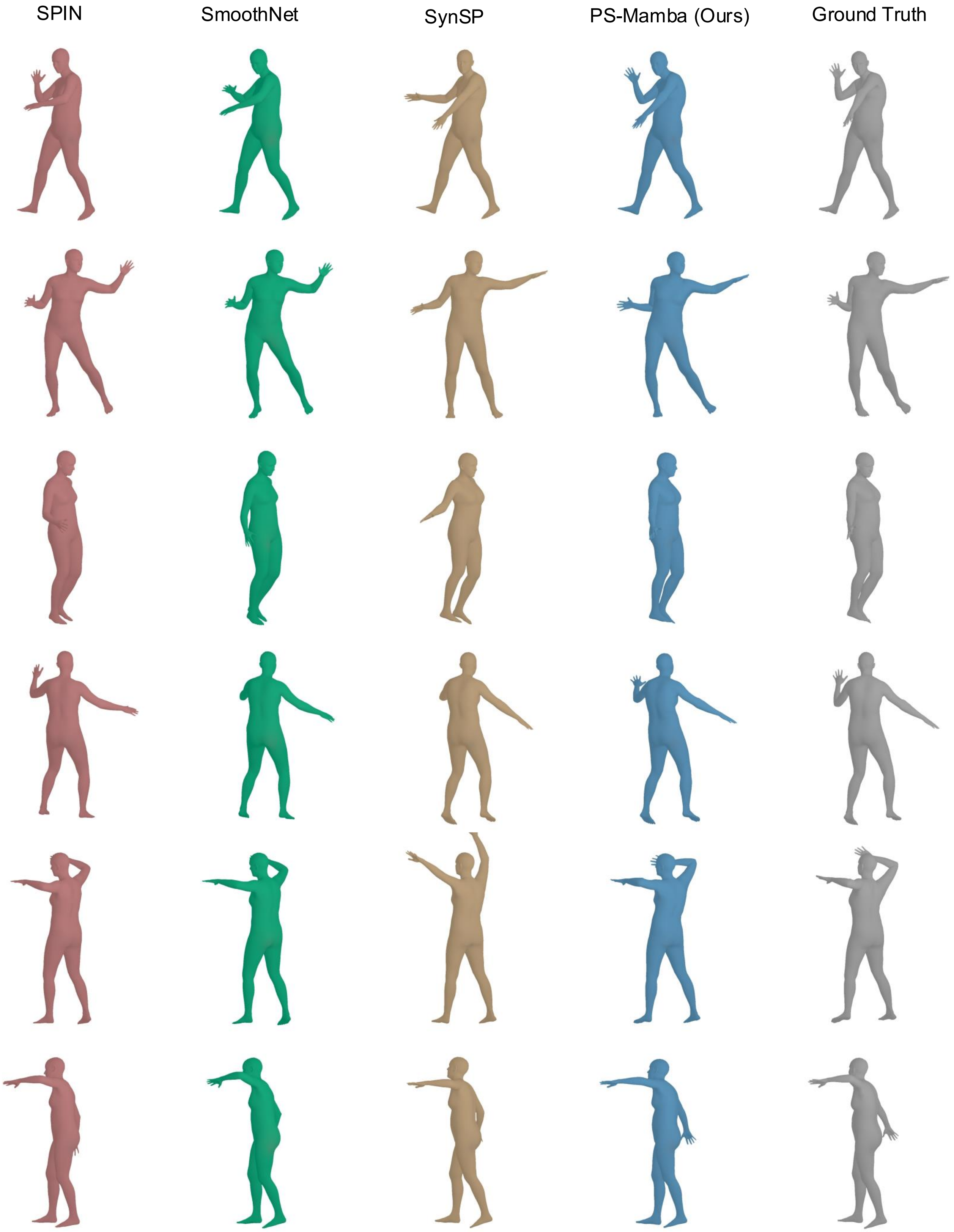

Figure 5. **Qualitative comparisons on AIST++ [12, 17] Dataset.** We compare our method with SynSP [20] and SmoothNet [23]. Our approach refines 3D meshes more accurately, particularly in ambiguous cases, demonstrating our PS-Mamba's effectiveness in handling uncertainty and ensuring robustness in challenging scenarios.

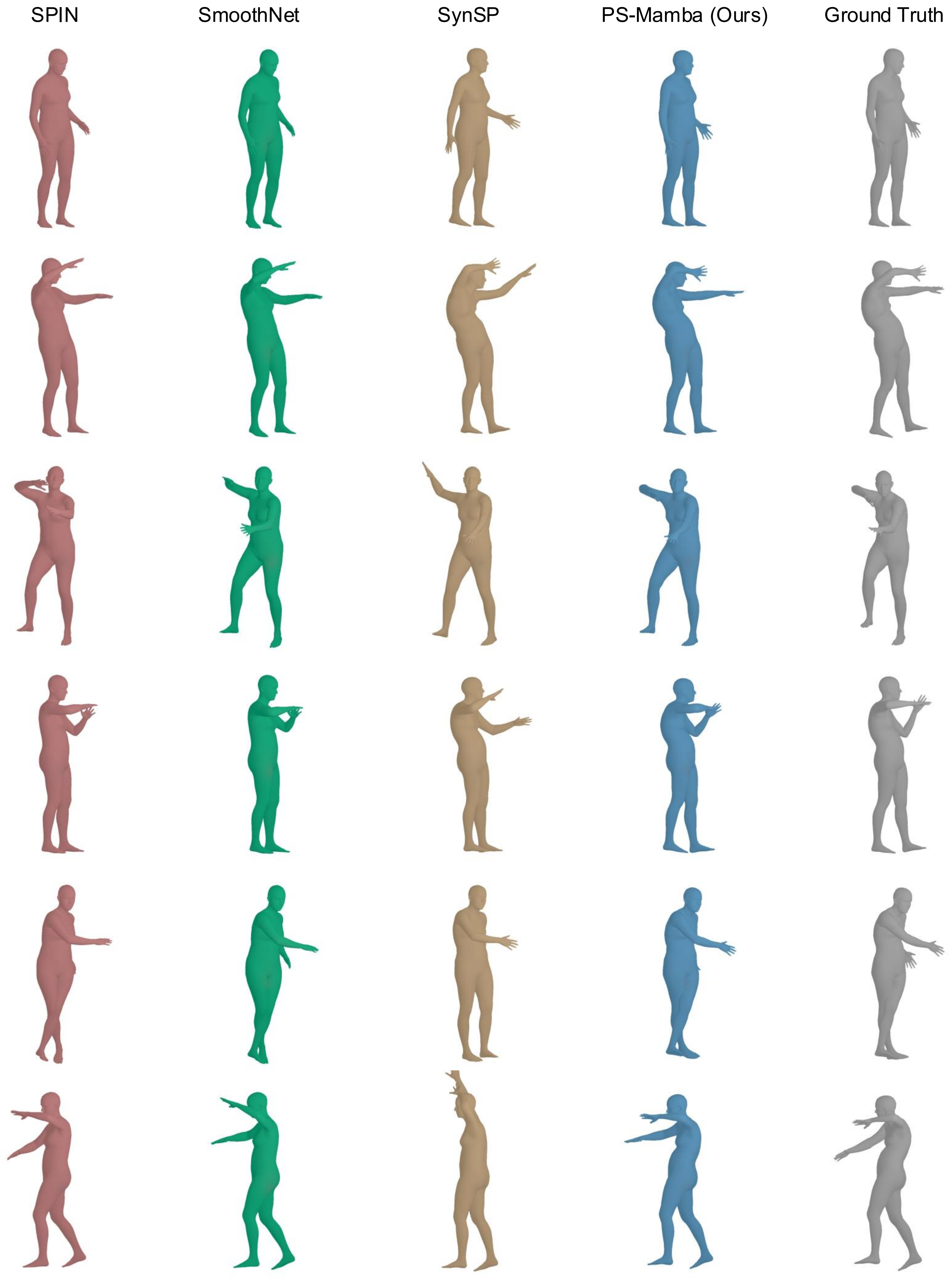

Figure 6. **Qualitative comparisons on AIST++ [12, 17] Dataset.** We compare our method with SynSP [20] and SmoothNet [23]. Our approach refines 3D meshes more accurately, particularly in ambiguous cases, demonstrating our PS-Mamba's effectiveness in handling uncertainty and ensuring robustness in challenging scenarios.

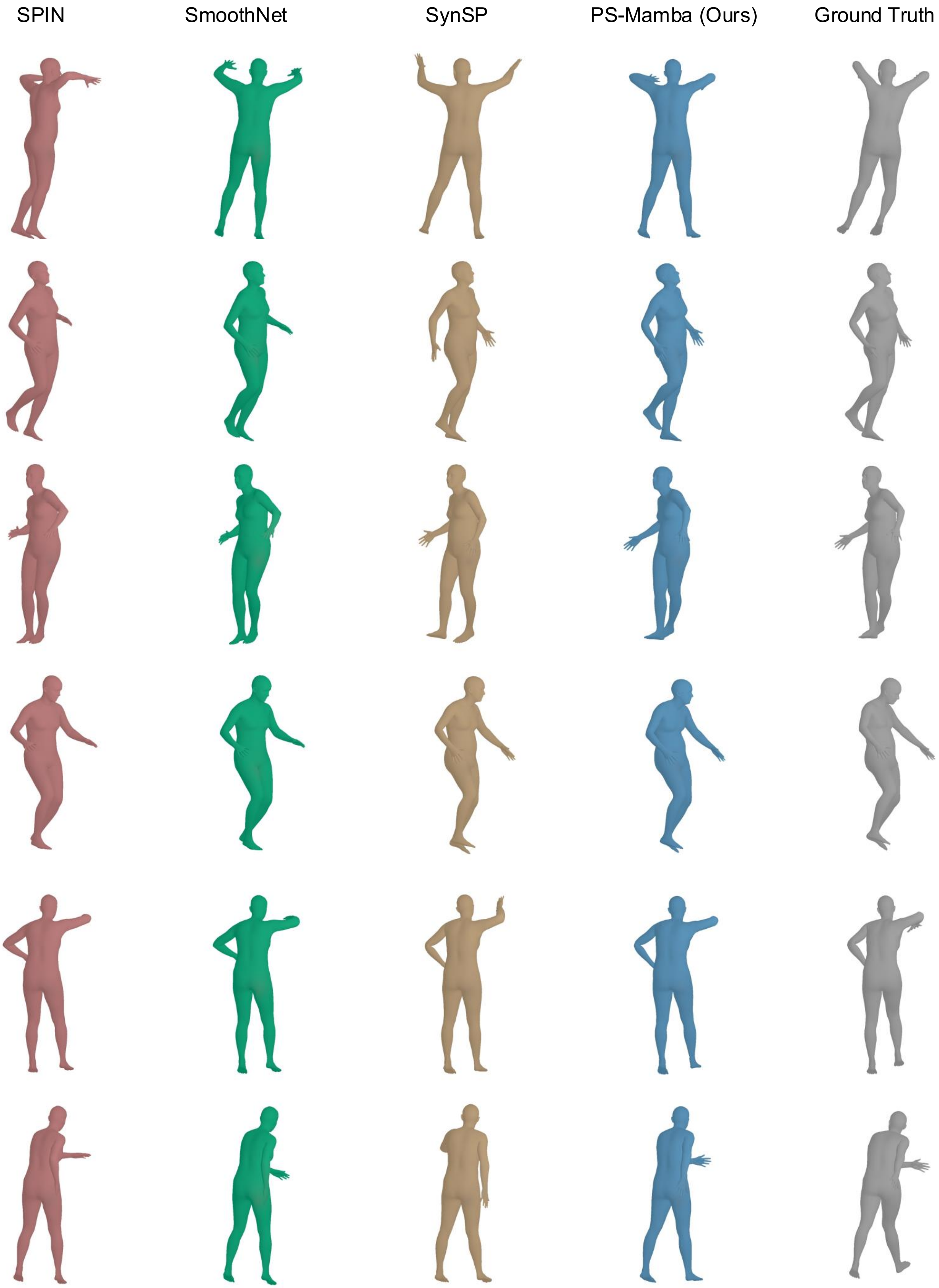

Figure 7. **Qualitative comparisons on AIST++ [12, 17] Dataset.** We compare our method with SynSP [20] and SmoothNet [23]. Our approach refines 3D meshes more accurately, particularly in ambiguous cases, demonstrating our PS-Mamba's effectiveness in handling uncertainty and ensuring robustness in challenging scenarios.

References

[1] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. PandaNet: Anchor-based single-shot multi-person 3D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6856–6865, 2020. 2, 3

[2] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ Filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2527–2530, 2012. 2

[3] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. GFPose: Learning 3D human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4810, 2023. 2

[4] Ziqiang Dang, Tianxing Fan, Boming Zhao, Xujie Shen, Lei Wang, Guofeng Zhang, and Zhaopeng Cui. MoManifold: Learning to measure 3D human motion via decoupled joint acceleration manifolds. In *Proceedings of British Machine Vision Conference (BMVC)*, 2024. 2

[5] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 3

[6] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3, 4, 5

[7] Kyung-Min Jin, Byoung-Sung Lim, Gun-Hee Lee, Tae-Kyung Kang, and Seong-Whan Lee. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5725–5734, 2023. 3

[8] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 2, 3

[9] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263, 2020. 2, 3, 5

[10] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 3

[11] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2

[12] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13401–13412, 2021. 2, 3, 5, 6, 7, 8

[13] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. VMamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1

[14] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2640–2649, 2017. 2, 3, 5

[15] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 2

[16] William H Press and Saul A Teukolsky. Savitzky-golay smoothing filters. *Computers in Physics*, 4(6):669–672, 1990. 2

[17] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST Dance Video Database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, pages 501–510, 2019. 2, 3, 5, 6, 7, 8

[18] Carnegie Mellon University. CMU Graphics Lab motion capture database. `http://mocap.cs.cmu.edu/`, 2003. 2

[19] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 2, 3

[20] Tao Wang, Lei Jin, Zheng Wang, Jianshu Li, Liang Li, Fang Zhao, Yu Cheng, Li Yuan, Li Zhou, Junliang Xing, et al. SynSP: Synergy of smoothness and precision in pose sequences refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1824–1833, 2024. 2, 3, 6, 7, 8

[21] Ian T Young and Lucas J Van Vliet. Recursive implementation of the gaussian filter. *Signal processing*, 44(2):139–151, 1995. 2

[22] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. DeciWatch: A simple baseline for 10x efficient 2D and 3D pose estimation. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 3

[23] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. SmoothNet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision (ECCV)*, pages 625–642. Springer, 2022. 2, 3, 6, 7, 8