KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation—Supplementary Material

Jihua Peng[1,2] Yanghong Zhou[1] P. Y. Mok[1,2,*]
[1]The Hong Kong Polytechnic University [2]Laboratory for Artificial Intelligence in Design
{ji-hua.peng, yanghong.zhou}@connect.polyu.hk, tracy.mok@polyu.edu.hk

The supplementary material contains: 1) adaptable to different 3D pose estimators; 2) ablation studies on Human3.6M; 3) more qualitative results.

1. Adaptable to Different 3D Pose Estimators

Implementation Details. In this section, we illustrate in detail on how our Kinematics Prior Attention (KPA) and Trajectory Prior Attention (TPA) are applied to different 3D pose estimators. Our TPA possesses the capability to not only model joint-to-joint motion trajectory across frames but also to model pose-to-pose motion trajectory across frames. Figure 1 shows the joint-to-joint and pose-to-pose motion trajectory topology. In Figure 1(b), TPA connects the different poses across consecutive adjacent frames to build the temporal local topology (pose-to-pose), including self-connection. Next, we exploit learnable vectors (dotted line) to connect the poses among all neighbouring and non-neighbouring frames to construct the simulated temporal global topology (pose-to-pose), which is equivalent to the computation of attention weights among all frames by the self-attention. Then, the two topologies are integrated together through the combination method identical to joint motion trajectory topology (Figure 1(a)), resulting in the pose motion trajectory topology. The pose motion trajectory topology (Figure 1(b)) is incorporated into the stacked TPA (pose) to encode the pose-to-pose features across frames for these works [2, 3, 7]. On the other hand, we introduce joint motion trajectory topology (Figure 1(a)) into the stacked TPA (joint) to learn joint-to-joint temporal information for other works [5, 6]. Figure 2 depicts the framework overview of our KPA and TPA applied to different 3D pose estimators. For PoseFormer [7], the KPA and the stacked TPA (pose) are placed ahead of the stacked spatial transformers and stacked temporal transformers, respectively. The model architecture of StridedTransformer [2] with our method is similar to PoseFormer [7]. Hence, we have not depicted it. For MHFormer [3], we employ the KPA to process the initial 2D pose sequence, generating Q, K and V vectors for the first spatial transformer. Then, we utilize three parallel stacked TPA (pose) blocks to encode the pose-to-pose temporal features for multiple hypotheses, respectively. The three outputs from three stacked TPA (pose) blocks are fed into the next layer. In terms of STCFormer [5], the KPA and the stacked TPA (joint) blocks are positioned ahead of the spatial attention and temporal attention in parallel. They yield spatial and temporal Q, K and V vectors with priori knowledge for the spatial attention and temporal attention, respectively. For D3DP [4], we employ two KPA blocks to concurrently process the 2D pose sequence and noisy 3D pose sequence, subsequently concatenating the output features and feeding them into the spatial transformer. Then, the stacked TPA (joint) blocks are placed between the spatial transformer and temporal transformer. D3DP [4] adopts the MixSTE [6] as the denoiser, so the model architecture of MixSTE [6] with our method is similar to D3DP [4].

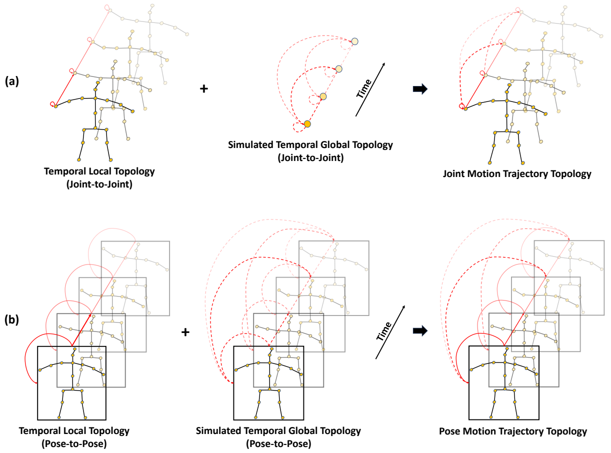

(a)
Temporal Local Topology
(Joint-to-Joint)
+
Simulated Temporal Global Topology
(Joint-to-Joint)
Time
Joint Motion Trajectory Topology
(b)
Temporal Local Topology
(Pose-to-Pose)
+
Simulated Temporal Global Topology
(Pose-to-Pose)
Time
Pose Motion Trajectory Topology

Figure 1. Overview of different motion trajectory topology. (a) The temporal local topology (joint-to-joint) plus the simulated temporal global topology (joint-to-joint) to form the joint motion trajectory topology. (b) The temporal local topology (pose-to-pose) plus the simulated temporal global topology (pose-to-pose) to form the pose motion trajectory topology.

Enhanced Attention Maps. In this section, we visual-

[*]Corresponding author

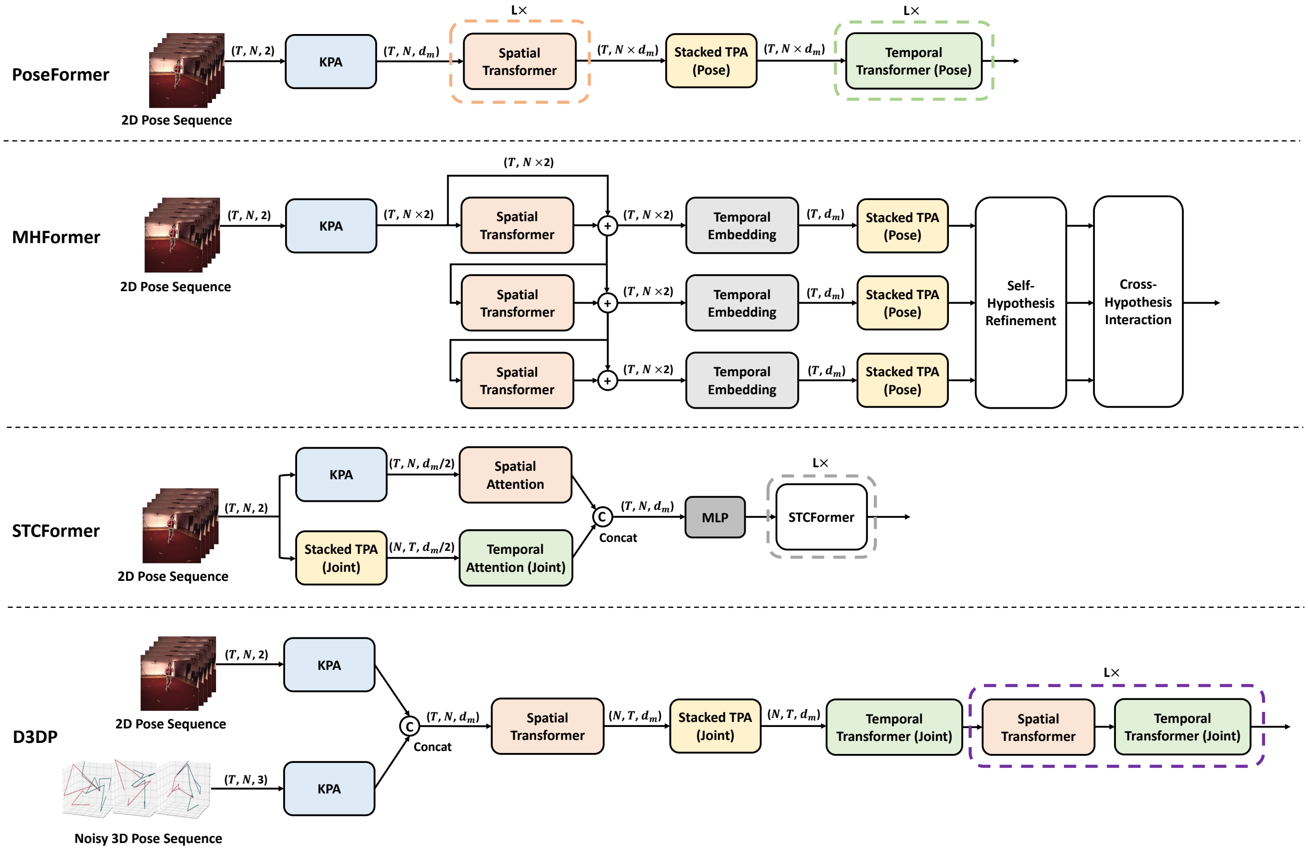


Figure 2. The framework overview of our KPA and TPA applied to different 3D pose estimators. The stacked TPA indicate that two TPA blocks are stacked with a residual connection. In terms of PoseFormer [7] and MHFormer [3], we use the stacked TPA (pose) to model temporal correlations between poses across frames. In contrast, the stacked TPA (joint) is utilized to encode the temporal features between joints across frames for STCFormer [5] and D3DP [4].

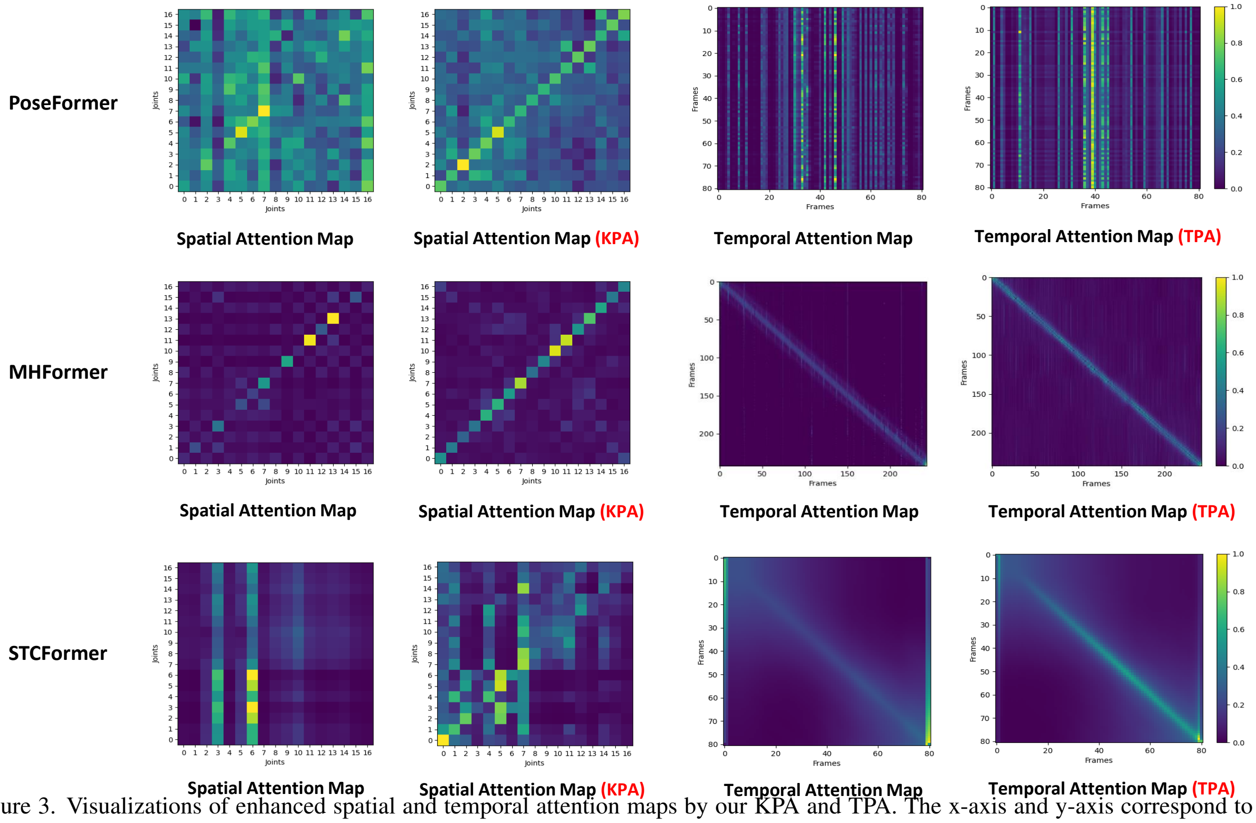
ize the enhanced attention maps of [3, 5, 7] after applying our KPA and TPA on Human3.6M, to validate the effectiveness of our method. Figure 3 illustrates enhanced spatial and temporal attention maps from PoseFormer [7], MHFormer [3] and STCFormer [5], by integrating our KPA and TPA into their networks. In terms of spatial attention maps, our KPA enhances attention weights between certain joints based on human anatomical structures and kinematic relationships, facilitating the explicit representation of human body topological relationships in the attention maps. On the other hand, our TPA enhances the temporal correlations between adjacent frames based on the motion trajectories of poses or joints in MHFormer [3] and STCFormer [5]. In particular, our TPA enhances attention weights between the frames of central region and other frames in PoseFormer [7], recognizing the periodic nature of human motion in videos.

2. Ablation Studies on Human3.6M

Different Numbers of Modules. In this section, we validate the impact of different numbers of KPA and TPA blocks in the KTPFormer. Table 1 reports the MPJPE and P-MPJPE comparisons on Human3.6M dataset. We take the estimated 2D poses by CPN [1] as input and train these models under 81 frames. The baseline network utilizes the stacked spatio-temporal encoders (L = 8) with number of heads H=8 and feature size C=512 to predict the 3D pose sequence. In our KTPFormer (first block), we combine KPA and TPA respectively with vanilla spatial transformer and temporal transformer, forming Kinematics-Enhanced Transformer and Trajectory-Enhanced Transformer, which are placed at the beginning of the network. Subsequently, we employ the stacked spatio-temporal encoders (L = 7) to encode features. In the KTPFormer (all blocks), we stack the Kinematics-Enhanced Transformer

Table 1. The MPJPE and P-MPJPE comparisons with different numbers of KPA and TPA blocks in the KTPFormer. The evaluation is performed on Human3.6M with 81 input frames. The best result in each column is marked in red.

Method	Parameters (M)	FLOPs (M)	MPJPE (mm)	P-MPJPE (mm)
Baseline	33.650	46346	43.1	34.1
KTPFormer (all blocks)	33.673	46412	42.3	33.4
KTPFormer (first block)	33.652	46353	41.8	32.6



Spatial Attention Map | Spatial Attention Map (KPA) | Temporal Attention Map | Temporal Attention Map (TPA)

Spatial Attention Map | Spatial Attention Map (KPA) | Temporal Attention Map | Temporal Attention Map (TPA)

Spatial Attention Map | Spatial Attention Map (KPA) | Temporal Attention Map | Temporal Attention Map (TPA)

Figure 3. Visualizations of enhanced spatial and temporal attention maps by our KPA and TPA. The x-axis and y-axis correspond to the queries and the predicted outputs, respectively. The attention weights are normalized from 0 to 1, and the lighter color indicates stronger attention.

and Trajectory-Enhanced Transformer for $L = 8$ loops. As indicated by the results, our KTPFormer (first block) obtains the lowest errors of MPJPE and P-MPJPE, indicating that KPA and TPA are better suited for processing the initial 2D pose sequence. Also, the KTPFormer (first block) can improve the performance more efficiently and has only a smaller increase in the computational overhead compared to the KTPFormer (all blocks). The design of KTPFormer (first block) is more effectively applicable to different 3D pose estimators.

Different Combination Ways of Topologies. We compare two different ways of combining the local topology and the simulated global topology. The first combination has been illustrated in the main text. We apply the first combination way to our KTPFormer, namely KTPFormer (average). The second combination is to directly add the local topology and the simulated global topology to obtain the kinematics topology or the joint motion trajectory topology. The second combination way is also applied to the KTPFormer, called KTPFormer (add). We train the two networks using the estimated 2D poses by CPN [1] with 81 frames as input. As shown in Table 2, the KTPFormer (average) achieves the best results of MPJPE and P-MPJPE. It suggests that the KTPFormer (average) which ensures the symmetry of the final topology allows the nodes to learn the spatial or temporal prior knowledge between them without being influenced by the direction of node connections.

Table 2. The MPJPE and P-MPJPE comparisons with different combination ways of topologies in the KPA and TPA. The evaluation is performed on Human3.6M with 81 input frames. The best result in each column is marked in red.

Method	Parameters (M)	FLOPs (M)	MPJPE (mm)	P-MPJPE (mm)
KTPFormer (add)	33.652	46353	42.1	33.3
KTPFormer (average)	33.652	46353	41.8	32.6

Free Parameters. We conduct experiments on the KTPFormer under three free parameters, including the number of spatio-temporal encoders L, the feature size of transformer layers C and the number of heads H, to examine different architectures of KTPFormer. During the experiment, we alter each free parameter while maintaining a constant value for the remaining two parameters. Table 3 reports the comparisons on Human3.6M using the CPN's 2D pose detection with 81 frames as input. The KTPFormer with L = 7, C = 512 and H = 8 achieves the runner-up result of MPJPE and the best result of P-MPJPE, and strikes a balance between regression capacity and computational cost. Thus, we choose this configuration as the standard version of KTPFormer.

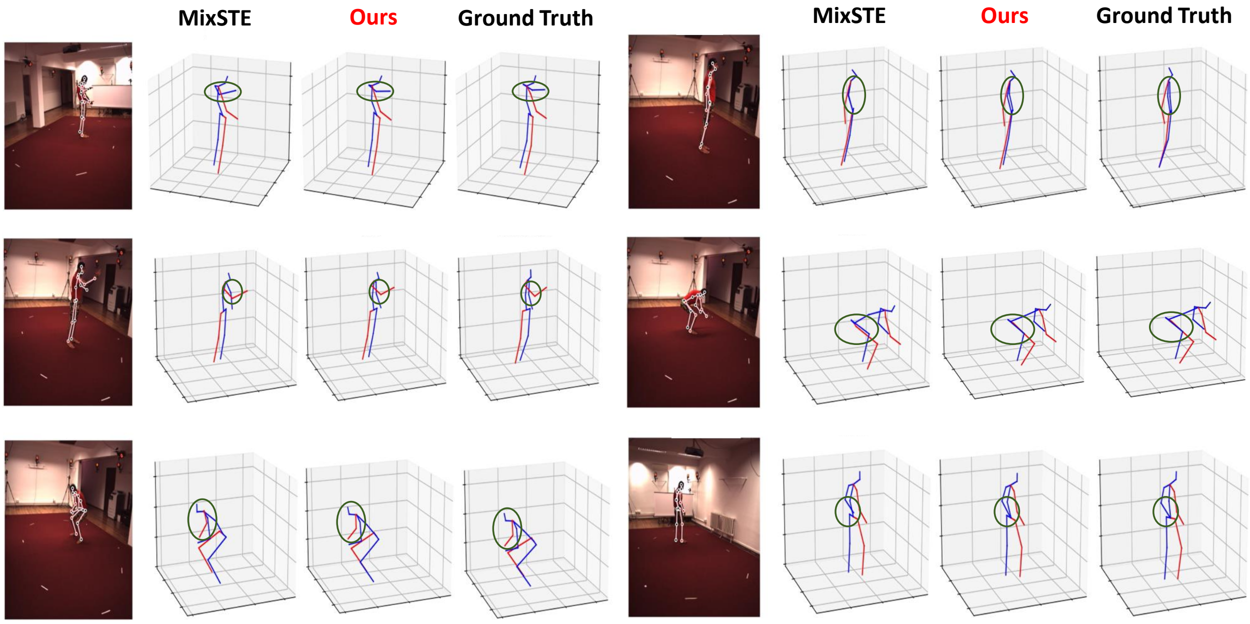


Figure 4. Visual comparisons of 3D pose estimation between MixSTE [6] and our KTPFormer on Human3.6M dataset. The green circle highlights locations where our KTPFormer yields better results.

Table 3. The MPJPE and P-MPJPE of KTPFormer with different number of spatio-temporal encoders L, feature size of transformer layers C, and the number of heads H in self-attention on Human3.6M dataset. Red: Best results. Blue: Runner-up results.

L	C	H	Parameters (M)	FLOPs (M)	MPJPE (mm)	P-MPJPE (mm)
6	512	4	29.446	40560	42.2	33.4
7	512	4	33.652	46353	41.7	33.1
8	512	4	37.857	52145	43.0	33.6
7	256	4	8.437	11625	43.0	33.7
7	512	4	33.652	46353	41.7	33.1
7	1024	4	134.413	185115	42.5	33.7
7	512	1	33.652	46353	43.0	34.2
7	512	2	33.652	46353	42.8	33.8
7	512	4	33.652	46353	41.7	33.1
7	512	8	33.652	46353	41.8	32.6
7	512	16	33.652	46353	42.5	33.4

3. More Qualitative Results

In this section, we present more qualitative results of KTPFormer. Figure 4 presents visual comparisons of 3D pose estimation results between our KTPFormer and MixSTE [6]. The green circle highlights locations where we can achieve more accurate 3D pose estimations compared to MixSTE [6]. Furthermore, we collect several in-the-wild videos as an additional real-world test to validate the generalization ability of our method. As shown in Figure 5, our method demonstrates remarkable robustness and accuracy across the majority of frames in the wild videos, especially in challenging scenarios with severe occlusion and extremely fast movements.

References

[1] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2, 3

[2] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2022. 1

[3] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 1, 2

[4] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. *arXiv preprint arXiv:2303.11579*, 2023. 1, 2

[5] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 1, 2

[6] 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. 1, 4

[7] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 1, 2

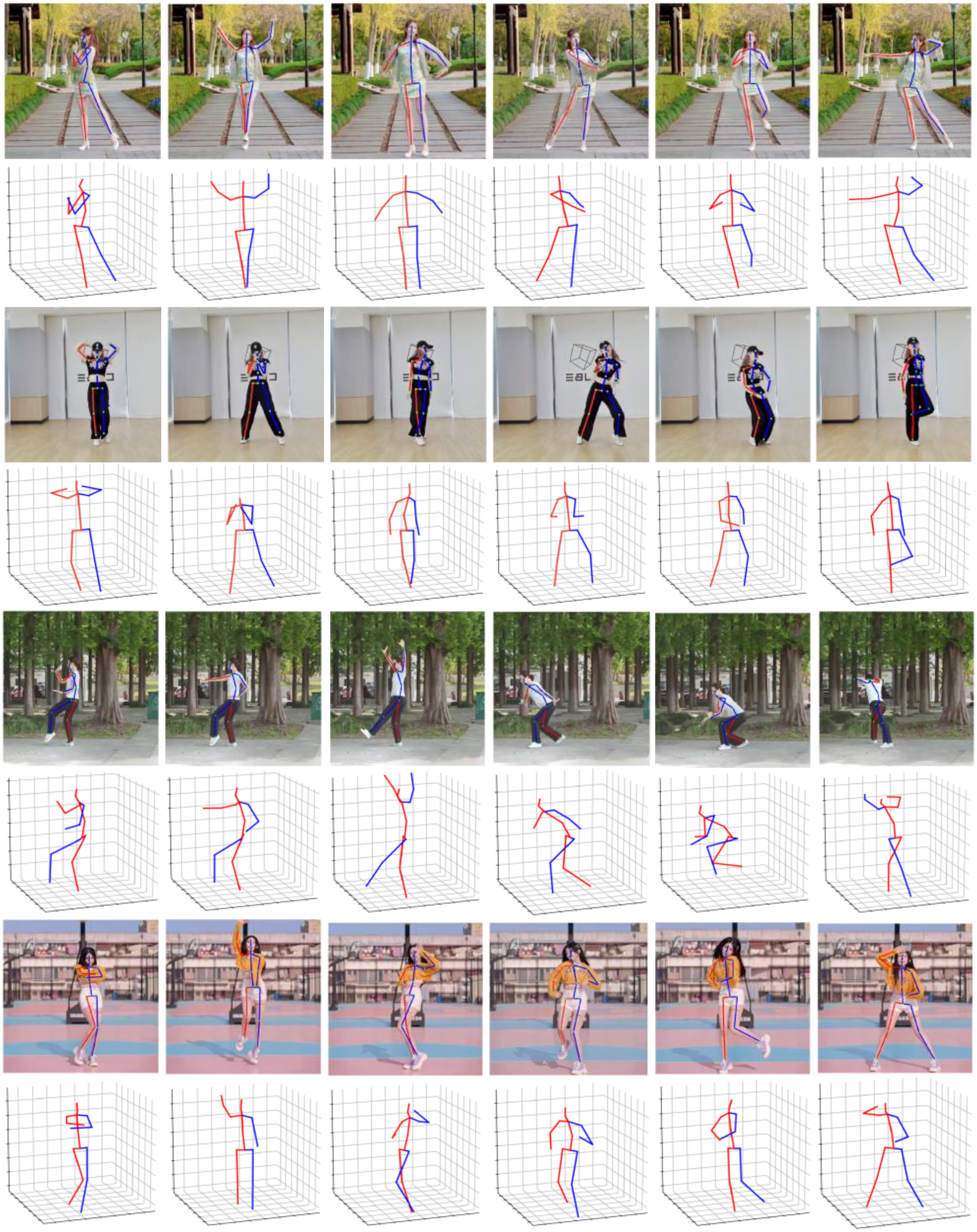


Figure 5. Some visualisation results of 3D pose estimation by our KTPFormer on in-the-wild videos.