3D Human Pose Estimation with Spatio-Temporal Criss-cross Attention — CVPR 2023 Supplementary Material*

Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, Ting Yao
Hefei University of Technology, Anhui, China HiDream.ai Inc
University of Science and Technology of China, Anhui, China
zhenhuat@foxmail.com, zhaofanqiu@gmail.com, haoyanbin@hotmail.com
hongrc.hfut@gmail.com, tingyao.ustc@gmail.com

Table 1. The P1 error comparisons with different attention modules on Human3.6M dataset. The best result in each column is marked in red.

Module	Frames T	Parameters	FLOPs (M)	P1(mm)
iterative	27	5.91M	2707	57.8
additive	27	4.72M	2166	66.3
STC	27	4.72M	2166	57.0

The supplementary material contains: 1) the ablation studies of attention module, positional embedding, post-processing, and free parameters on Human3.6M; 2) the per-joint error comparison on Human3.6M; 3) more qualitative analyses; 4) the code release of our implementations.

1. Ablation Studies on Human3.6M

1.1. Attention Module

To verify the effectiveness of our proposed STC block, we compare two attention block variants, i.e., iterative attention and additive attention, with different decomposition strategies of spatial and temporal attentions. Figure 1 shows the schematic illustration of the block variants. Specifically, **iterative** attention stacks a spatial attention layer and a temporal attention layer to simulate the full spatio-temporal attention. **Additive** attention also models spatial and temporal contexts in parallel but fuses the outputs of both attentions by element-wise summation. Table 1 summarizes the P1 error comparisons on Human3.6M dataset. Here we do not exploit the positional embedding for simplicity and take the estimated 27-frame 2D poses by CPN as input. As indicated by the results, our STC block obtains the lowest error and reaches a good trade-off between model capacity and computational cost.

*This work is supported by the National Natural Science Foundation of China under Grants 61932009.

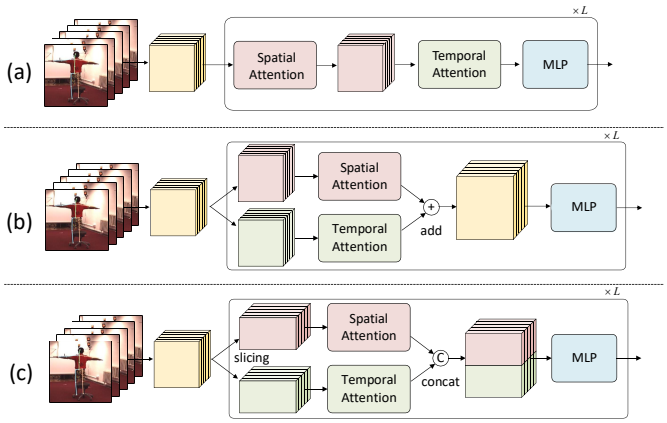

Figure 1. Modeling spatio-temporal correlation for 3D human pose estimation by (a) iterative capturing spatial and temporal context, (b) adding the outputs from two separate attention layers, and (c) our Spatio-Temporal Criss-cross attention (STC), i.e., a two-pathway block that models spatial and temporal information in parallel and concatenates the output from attention layer.

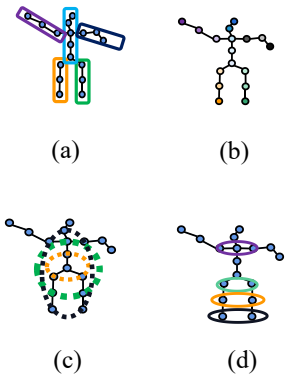

Figure 2. The illustration of different positional embedding functions: (a) SPE_1, (b) APE, (3) CPE, and (d) SyPE.

1.2. Positional Embedding

Next, we compare our SPE_1 with three different positional embedding functions, including Absolute Position

Table 2. The P1 error comparisons of different positional embedding functions on Human3.6M dataset. The best result in each column is marked in red.

		P1 (mm)
SPE_1	#1	48.3
APE	#2	48.9
CPE	#3	49.6
SyPE	#4	49.2

Embedding (APE), Centrality Positional Embedding (CPE) and Symmetric Positional Embedding (SyPE), by different separations of body joint groups, as shown in Figure 2.

Absolute positional embedding considers each joint independently and replaces the group index in SPE_1 with joint index. **Centrality positional embedding** divides the N body joints into six groups according to the adjacent distance to the hip joint. The groups in CPE are defined as

$$\begin{aligned}
g_0 &= \{hip\} \\
g_1 &= \{spine, right_hip, left_hip\} \\
g_2 &= \{thorax, right_knee, left_knee\} \\
g_3 &= \{neck, right_feet, left_feet, right_shoulder, left_shoulder\ \} \\
g_4 &= \{head, right_elbow, left_elbow\} \\
g_5 &= \{right_wrist, left_wrist\}
\end{aligned} \tag{1}$$

Instead, **symmetric positional embedding** divides N body joints into eleven groups according to the symmetrical structure of human body. The groups in SyPE are

$$\begin{aligned}
g_0 &= \{hip\} \\
g_1 &= \{spine\} \\
g_2 &= \{thorax\} \\
g_3 &= \{neck\} \\
g_4 &= \{head\} \\
g_5 &= \{right_hip, left_hip\} \\
g_6 &= \{right_knee, left_knee\} \\
g_7 &= \{right_feet, left_feet\} \\
g_8 &= \{right_shoulder, left_shoulder\} \\
g_9 &= \{right_elbow, left_elbow\} \\
g_{10} &= \{right_wrist, left_wrist\}
\end{aligned} \tag{2}$$

For each positional embedding function, the joints in the same group are attached with the same embedding vector. Table 2 compares our SPE_1 with the three positional embedding functions on Human3.6M dataset. In this experiment, we take the estimated 2D poses by CPN with 9 frames as input. Particularly, our SPE_1 achieves the lowest P1 error among four embedding functions, validating the advances of group separation by the dynamic chain structure in SPE_1.

Table 3. The P1 error comparisons with the state-of-the-art methods using post-processing on Human3.6M dataset. The best result in each column is marked in red.

Method	Publication	post-processing	**P1** ↓	
			CPN	GT
ST-GCN [1]	ICCV'19	✓	48.8	-
UGCN [6]	ECCV'20	✓	44.5	-
Einfalt [2]	aXiv'22	✓	44.2	-
StridedFormer [3]	TMM'22	✓	43.7	28.5
MHFormer [4]	CVPR'22	✓	42.4	-
P-STMO [4]	ECCV'22	✓	42.1	-
STCFormer		✓	40.8	21.3

1.3. Post-Processing

Recently, several works [1–6] employ the post-processing module proposed in [1] to improve the estimation accuracy. Accordingly, we further exploit the same post-processing in our STCFormer and compare with these baselines. Here, the models take the 243-frame estimated 2D poses by CPN or the 2D ground truth as input. As shown in Table 3, STCFormer achieves the best P1 error of both CPN input (40.8mm) and GT input (21.3mm).

1.4. Free Parameters

There are three free parameters for the STCFormer (i.e., the number of blocks L, the channel dimension C and the number of heads H). In this set of experiments, we test different values of these parameters to examine different architectures of STCFormer. In our implementations, we adjust each free parameter in order while fixing the other two parameters. Table 4 lists the comparisons. As indicated by the results, STCFormer with $L = 6$, $C = 256$ and $H = 8$ obtains the lowest error, and manages a good tradeoff between regression capacity and computational cost, that is regarded as the standard version of STCFormer.

Table 4. The P1 error of STCFormer with different number of blocks L, channel dimension of joint-based embedding C, and number of heads H in attention blocks on Human3.6M dataset. The error of default setting is marked in red.

L	C	H	Parameters	FLOPS (M)	P1(mm)
4	64	4	0.2M	91	50.0
6	64	4	0.3M	137	46.2
8	64	4	0.4M	183	47.5
6	128	4	1.19M	545	45.6
6	256	4	4.75M	2173	44.4
6	512	4	18.91M	8678	44.8
6	256	1	4.75M	2173	45.3
6	256	2	4.75M	2173	45.0
6	256	4	4.75M	2173	44.4
6	256	8	4.75M	2173	44.1
6	256	16	4.75M	2173	44.2

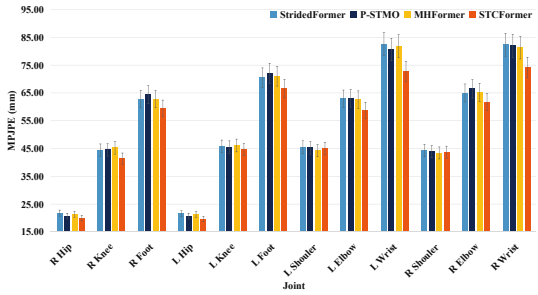

Figure 3. The per-joint error comparisons in terms of P1 with the state-of-the-art methods on Human3.6M dataset. 'L' and 'R' denote the left part and right part of the human body, respectively.

2. Per-joint Error Comparison on Human3.6M

In Figure 3, we compare the per-joint error of STCFormer and baseline methods. The input is 27-frame 2D poses estimated by CPN. A general performance tendency is observed that the joint errors increase along the body limbs, e.g., shoulder<elbow<wrist. Among the competitive methods including StridedFormer [3], P-STMO [5], and MHFormer [4], our STCFormer exhibits the best results on 10 out of 12 joints.

3. Inference Speed on Human3.6M

Table 5 here summarizes the P1 error and the inference speed on Human3.6M dataset. We measure the speed on a single P40 GPU, and compare two recent transformer-based methods of MHFormer [4] and MixSTE [7]. Overall, our STCFormer achieves the lowest P1 error and the fastest inference speed for both 27-frame and 81-frame inputs. For the 243-frame inputs, STCFormer also shows better trade-off than MHFormer and MixSTE. The results demonstrate the advantage of STC attention to decompose full spatio-temporal attention in an economic and effective way.

Table 5. The P1 error and inference speed on Human3.6M dataset. The 2D pose input is estimated by CPN. We measure the speed on a single P40 GPU. The best score is marked in red.

Method	Frames T	Speed (clip/s)	P1(mm)
MHFormer [4]	27	44	45.9
MixSTE [7]	27	46	45.1
STCFormer	27	72	44.1
MHFormer [4]	81	43	44.5
MixSTE [7]	81	35	42.7
STCFormer	81	65	42.0
MHFormer [4]	243	40	43.2
MixSTE [7]	243	12	40.9
STCFormer	243	40	41.0
STCFormer-L	243	21	40.5

4. Qualitative Analysis

In this section, we present more qualitative analyses of our STCFormer. Figure 4 and Figure 5 show more examples of visualized spatial attention and temporal attention, respectively, on Human3.6M dataset. Moreover, Figure 6 showcases 3D human pose estimation results by STCFormer and MHFormer on MPI-INF-3DHP dataset. In addition, to validate the generalization ability of our model, we crawled several videos from video Website as an additional real test and the pose estimation results by our STCFormer are illustrated in Figure 7. We also provide one video demo (demo.mp4), demonstrating the results of pose estimation on in-the-wild videos.

References

[1] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, 2019. 2

[2] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. *arXiv preprint arXiv:2210.06110*, 2022. 2

[3] 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*, 2022. 2, 3

[4] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, 2022. 2, 3, 5

[5] 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. *arXiv preprint arXiv:2203.07628*, 2022. 2, 3

[6] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *ECCV*, 2020. 2

[7] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. *arXiv preprint arXiv:2203.00859*, 2022. 3

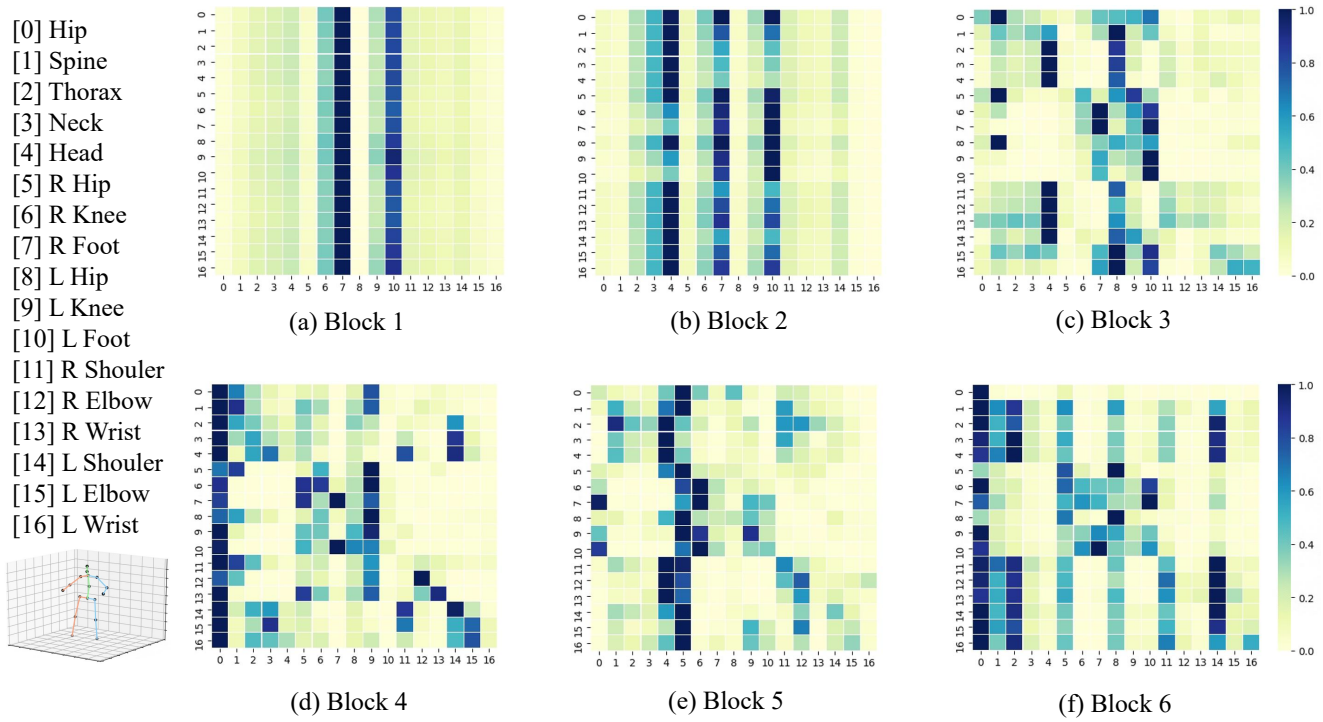

Figure 4. Visualizations of attention maps from the spatial attention modules in STCFormer. The x-axis and y-axis correspond to the queries and the predicted outputs, respectively.

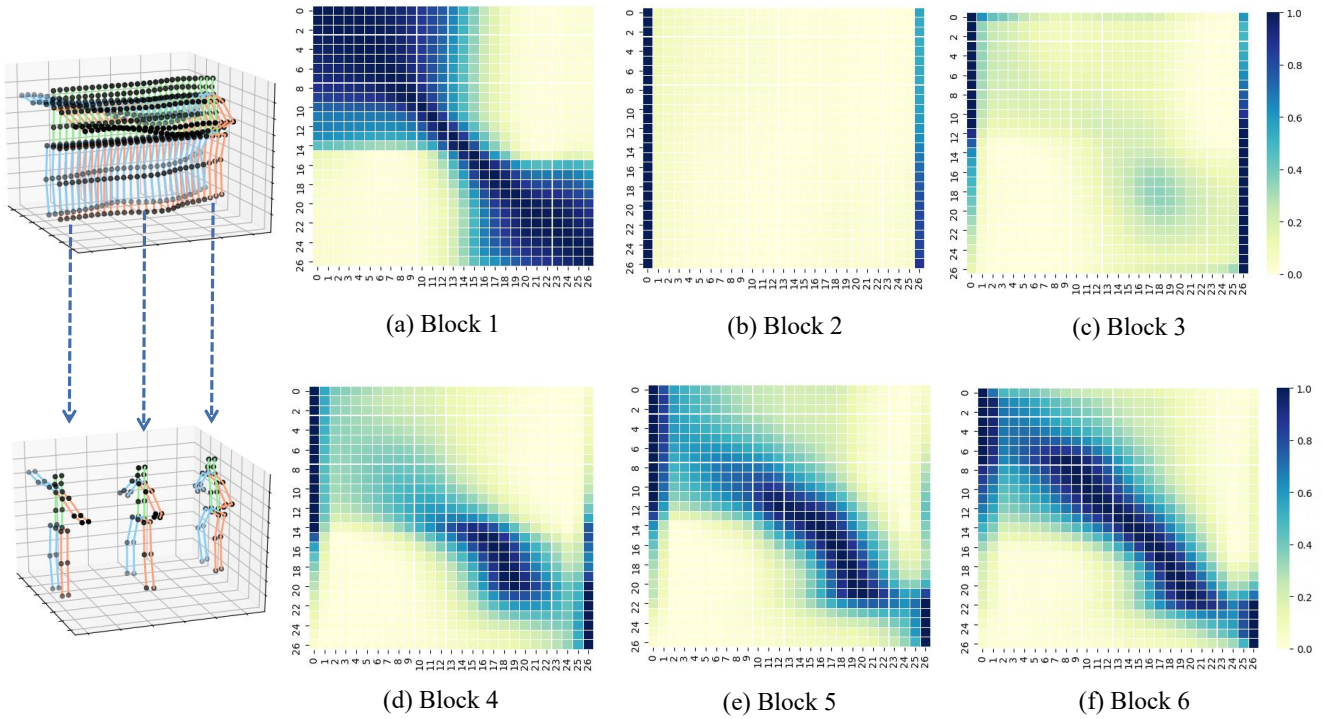

Figure 5. Visualizations of attention maps from the temporal attention modules in STCFormer. The x-axis and y-axis correspond to the queries and the predicted outputs, respectively.

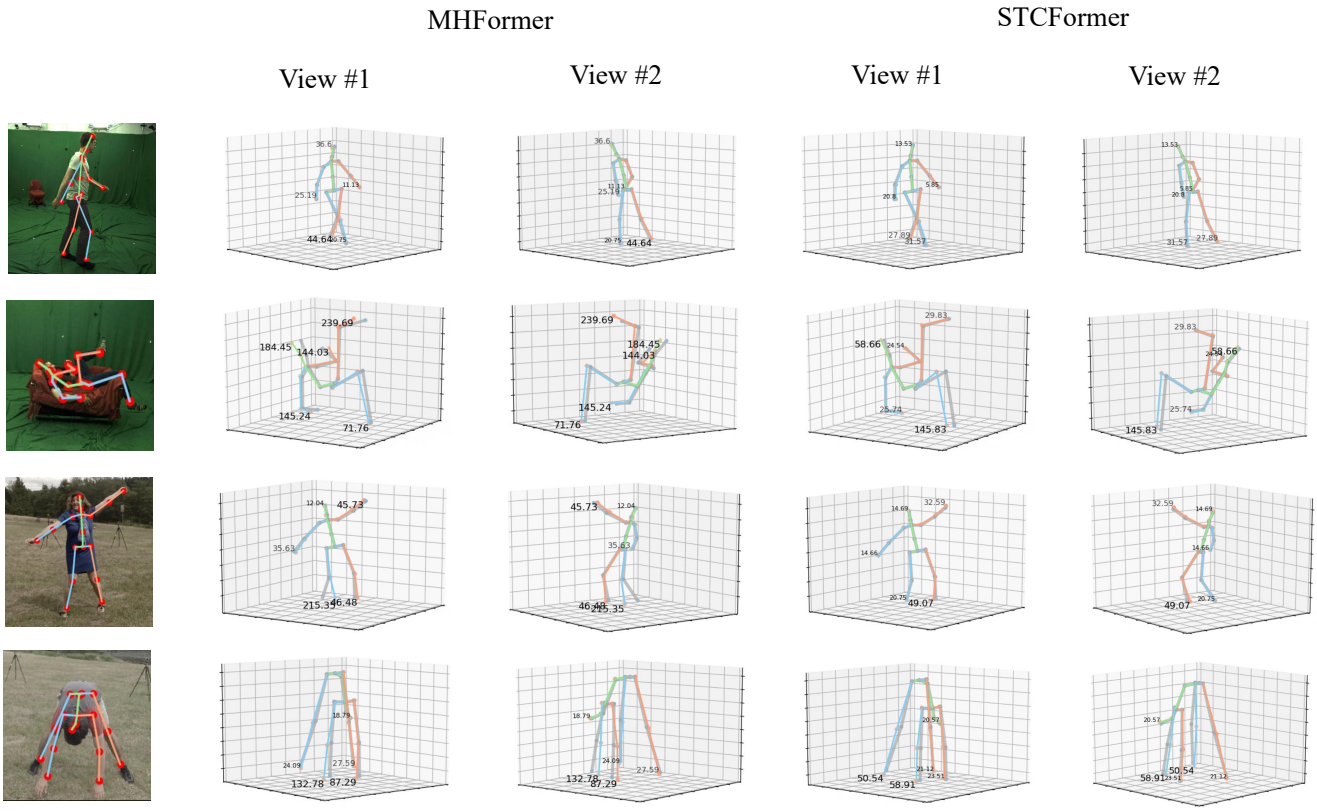

Figure 6. Examples of 3D pose estimation by MHFormer [4] and our STCFormer on MPI-INF-3DHP. The gray skeleton is the ground-truth 3D pose. Blue, orange and green skeletons represent the left part, right part and torso of the estimated human body, respectively. The number refers to the P1 error (mm) of joints in figure.

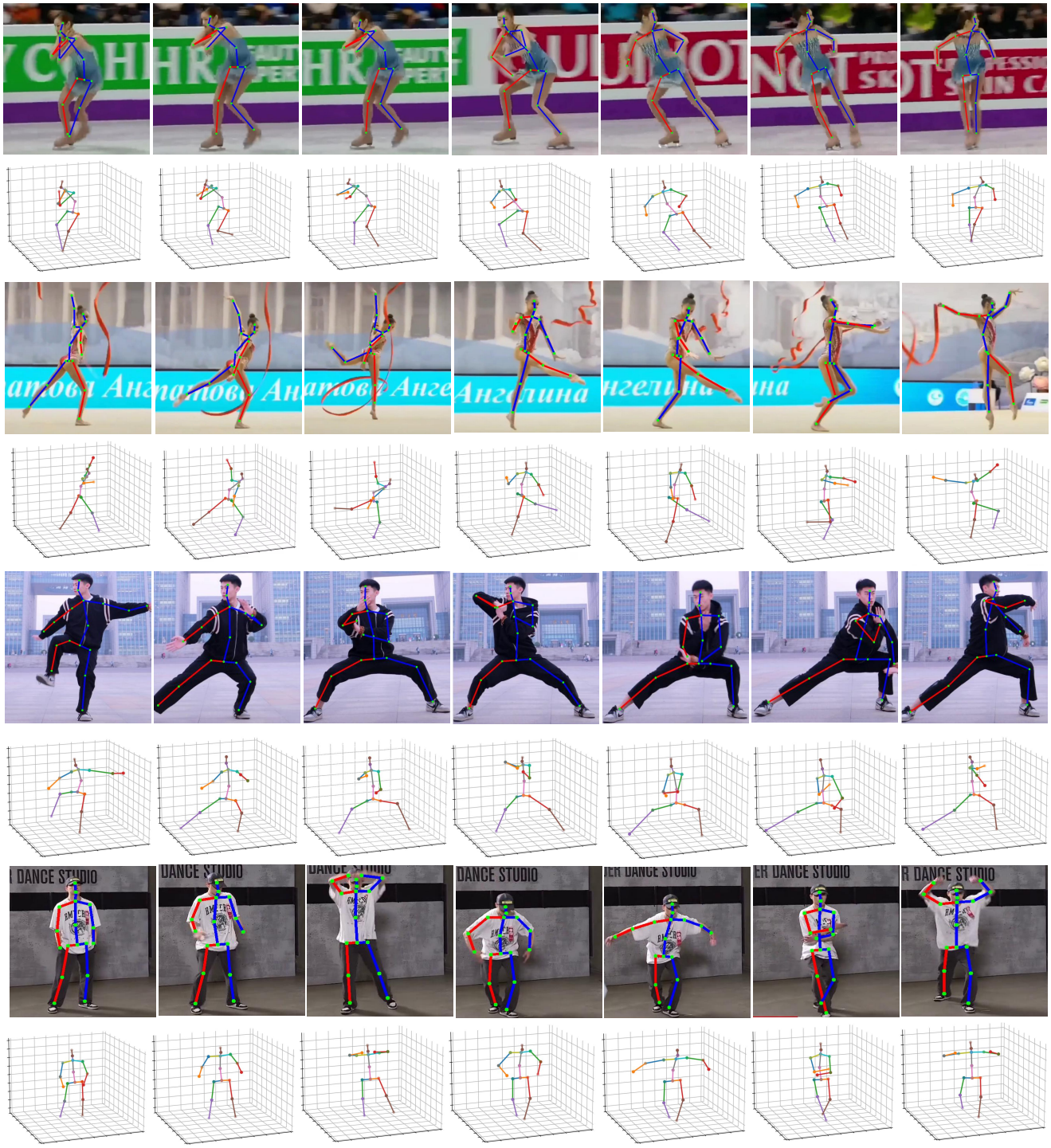

Figure 7. Examples of 3D pose estimation by our STCFormer on in-the-wild videos.