JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

TUE-PM-062

3D Human Pose Estimation with Spatio-Temporal Criss-cross Attention

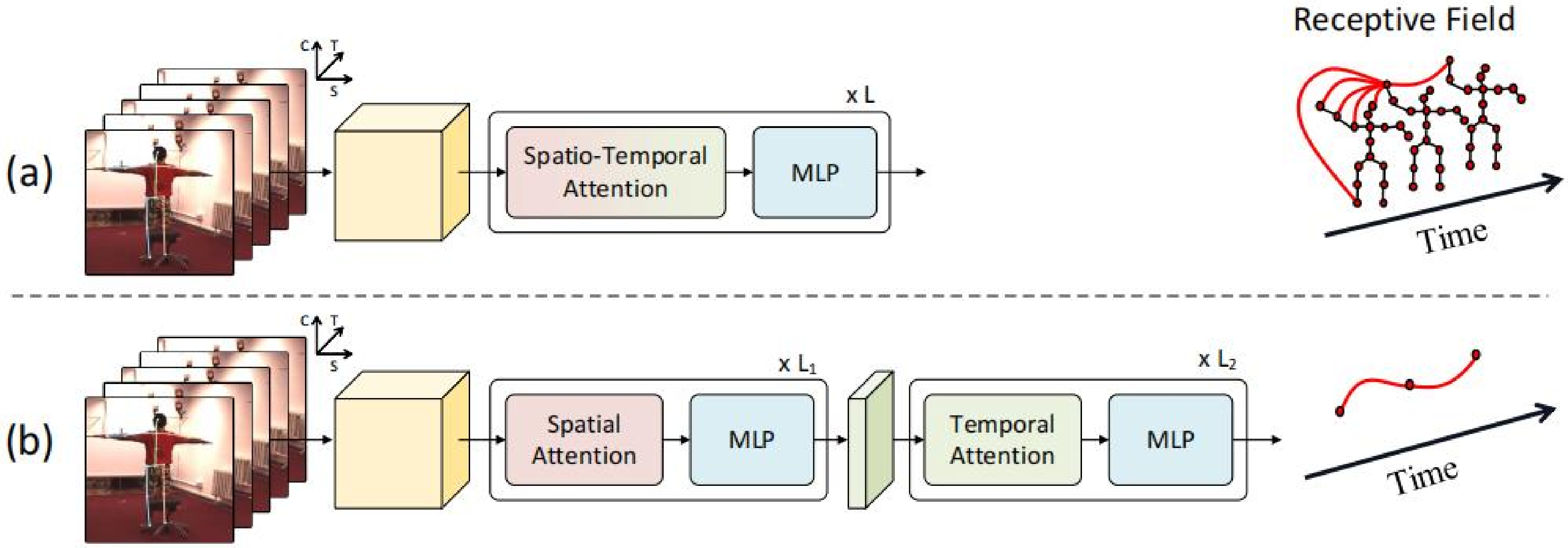
Zhenhua Tang , Zhaofan Qiu, Yanbin Hao, Richang Hong, Ting Yao

Hefei University of Technology
HiDream.ai Inc
University of Science and Technology of China

合肥工業大學
HEFEI UNIVERSITY OF TECHNOLOGY

中国科学技术大学
University of Science and Technology of China

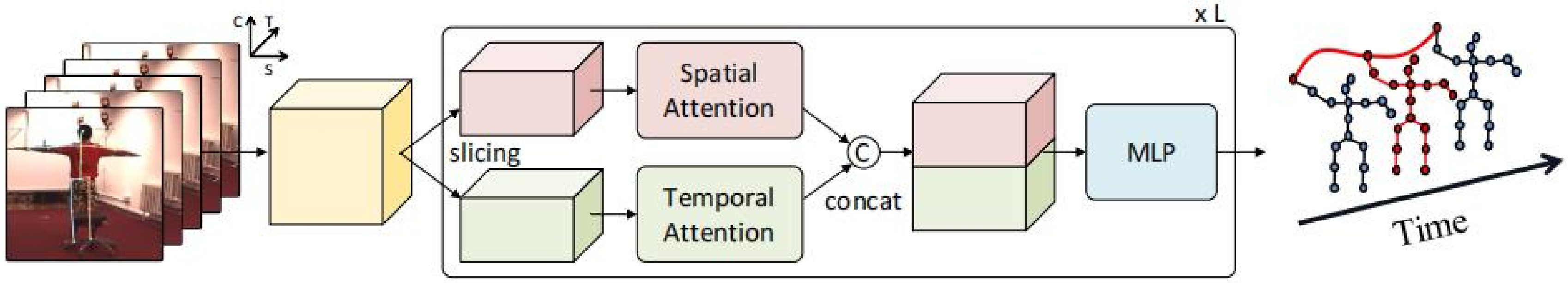
Spatio-Temporal Correlation for 3D Human Pose Estimation



(a) It has expensive computational cost with the increasing number of joints.
(b) It seldom explores the relation between joints across different frames.

Goals & Contributions

(c)



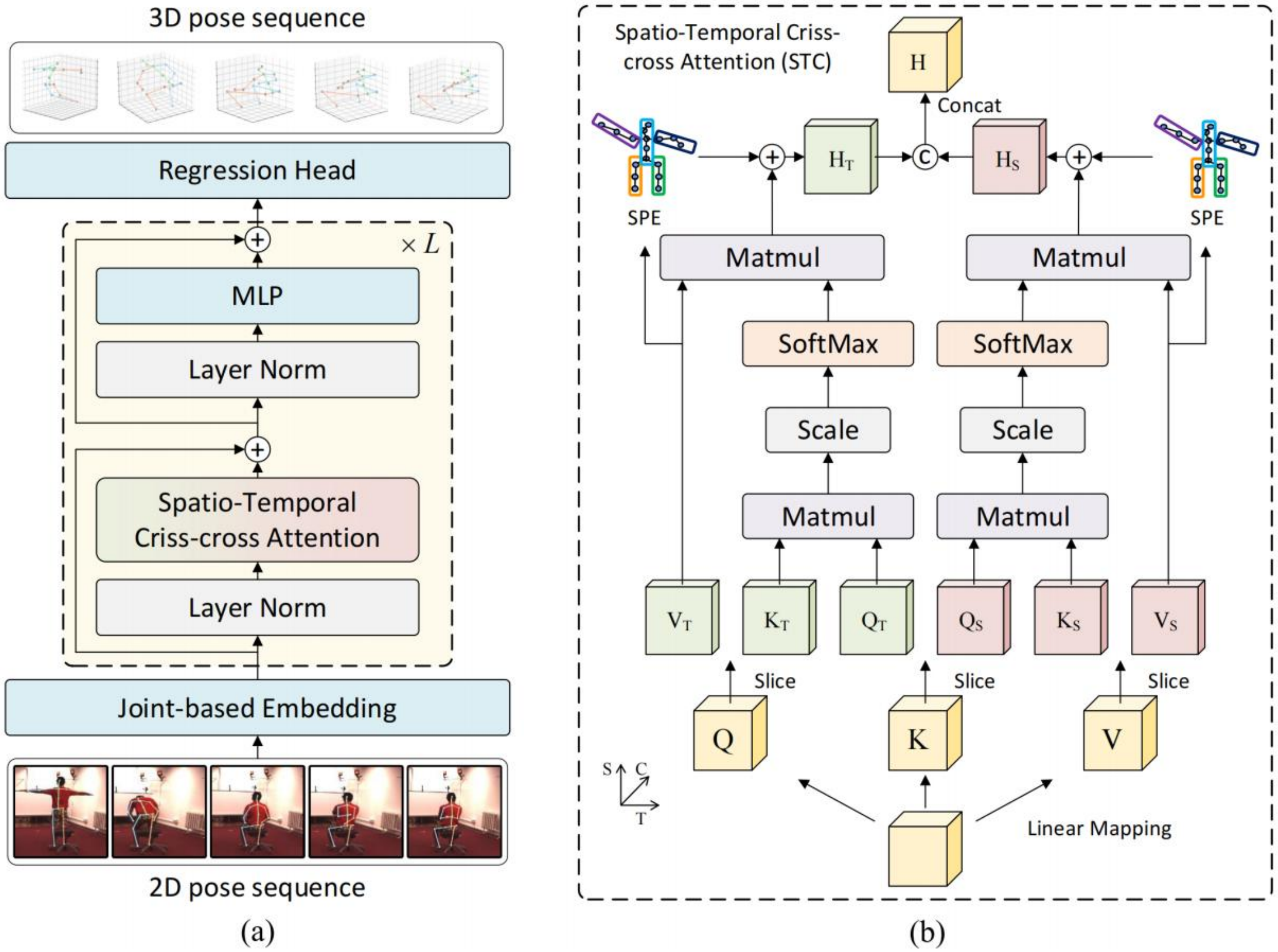
Goal:

- ✓ Modeling spatio-temporal correlation between joints spanning over the entire video with **minimal** computational overhead.

Contributions:

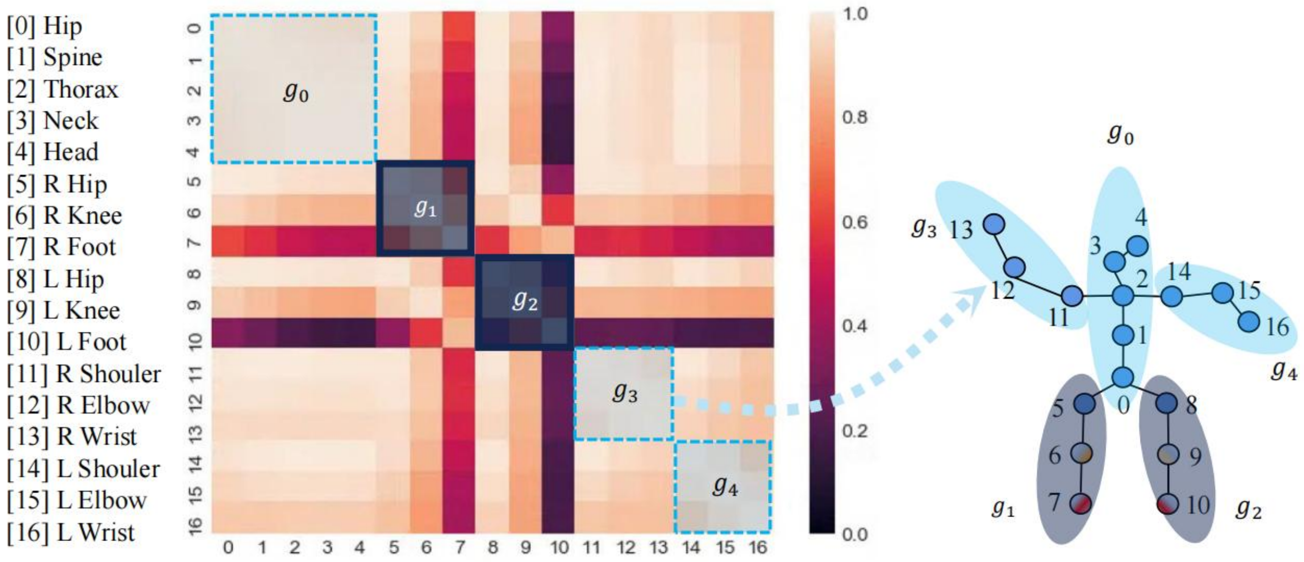
- ✓ We propose a novel two-pathway attention mechanism, namely **Spatio-Temporal Criss-cross attention(STC)**, that models the interactions between joints in an identical **frame** and joints in an identical **trajectory in parallel**.
- ✓ We devise STCFormer by stacking multiple STC blocks and further integrate a new **Structure-enhanced Positional Embedding (SPE)** into STCFormer to take the **structure** of **human body** into consideration.
- ✓ The proposed STCFormer with much less parameters **(-43.73.3%/43.65%** extra parameters/FLOPs) achieves superior performances than the state-of-the-art techniques **(-0.5 mm MPJPE) on Human3.6M**.

Workflow of Spatio-Temporal Criss-cross Transformer (STCFormer)



The **space pathway** computes the affinity between joints in each frame independently, and the **time pathway** correlates the identical joint moving across different frames, i.e., the trajectory.

Structure-enhanced Positional Embedding (SPE)



- The trajectories of joints in the **static part (g0, g3 and g4 in the figure)** are highly relevant.

- In the **dynamic part**, i.e., **part with relative movements (g1, g2 in the figure)**, the trajectories of joints are not relevant.

Performance Comparison on Human3.6M

Table 1. Performance comparisons in terms of P1 error (mm) and P2 error (mm) with the state-of-the-art methods on Human3.6M dataset. The 2D pose input is estimated by CPN [7]. The best result and runner-up result in each column are marked in red and blue, respectively. "*" denotes the post-processing module proposed in [4]. T is the number of sampled frames from each video.

P1	Publication	Dir.	Dis.	Eat.	Gre.	Phone	Photo	Pose	Purch.	Sit.	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Liu *et al.* [26] (T=243)	CVPR'20	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
UGCN [46] (T=96) *	ECCV'20	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
PoseFormer [54] (T=81)	ICCV'21	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
Shan *et al.* [40] (T=243)	ACM MM'21	40.8	44.5	41.4	42.7	46.3	55.6	41.8	41.9	53.7	60.8	45.0	41.5	44.8	30.8	31.9	44.3
Anatomy3D [6] (T=243)	TCVST'21	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
Einfalt *et al.* [9] (T=351) *	arXiv'22	39.6	43.8	40.2	42.4	46.5	53.9	42.3	42.5	55.7	62.3	45.1	43.0	44.7	30.1	30.8	44.2
StridedFormer [22] (T=243) *	TMM'22	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
CrossFormer [13] (T=81)	arXiv'22	40.7	44.1	40.8	41.5	45.8	52.8	41.2	40.8	55.3	61.9	44.9	41.8	44.6	29.2	31.1	43.7
PATA [48] (T=243)	TIP'22	39.9	42.7	40.3	42.3	45.0	52.8	40.4	39.3	56.9	61.2	44.1	41.3	42.8	28.4	29.3	43.1
MHFormer [23] (T=351)	CVPR'22	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
P-STMO [39] (T=243)	ECCV'22	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
MixSTE [52] (T=81)	CVPR'22	39.8	43.0	38.6	40.1	43.4	50.6	40.6	41.4	52.2	56.7	43.8	40.8	43.9	29.4	30.3	42.4
MixSTE [52] (T=243)	CVPR'22	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
STCFormer (T=81)		40.6	43.0	38.3	40.2	43.5	52.6	40.3	40.1	51.8	57.7	42.8	39.8	42.3	28.0	29.5	42.0
STCFormer (T=243)		39.6	41.6	37.4	38.8	43.1	51.1	39.1	39.7	51.4	57.4	41.8	38.5	40.7	27.1	28.6	41.0
STCFormer-L (T=243)		38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5

STCFormer-L outperforms StridedFormer , PATA and MixSTE with T=243 frames, which are also based on transformer architecture, by the P1 error drop of **3.2mm**, **2.6mm** and **0.4mm**, respectively.

Performance Comparison on Human3.6M

Method	Frames T	Parameters	FLOPs (M)	P1(mm)
StridedFormer [22]	27	4.01M	163	46.9
P-STMO [39]	27	4.6M	164	46.1
MHFormer [23]	27	18.92M	1000	45.9
MixSTE [52]	27	33.61M	15402	45.1
STCFormer	27	4.75M	2173	44.1
StridedFormer [22]	81	4.06M	392	45.4
P-STMO [39]	81	5.4M	493	44.1
MHFormer [23]	81	19.67M	1561	44.5
MixSTE [52]	81	33.61M	46208	42.7
STCFormer	81	4.75M	6520	42.0
StridedFormer [22]	243	4.23M	1372	44.0
P-STMO [39]	243	6.7M	1737	42.8
MHFormer [23]	243	24.72M	4812	43.2
MixSTE [52]	243	33.61M	138623	40.9
STCFormer	243	4.75M	19561	41.0
STCFormer-L	243	18.91M	78107	40.5

Method	Frames T	FPS	P1(mm)
MHFormer [2]	27	44	45.9
MixSTE [6]	27	46	45.1
STCFormer	27	72	44.1
MHFormer [2]	81	43	44.5
MixSTE [6]	81	35	42.7
STCFormer	81	65	42.0
MHFormer [2]	243	40	43.2
MixSTE [6]	243	12	40.9
STCFormer	243	40	41.0
STCFormer_L	243	21	40.5

The results again confirm the advances of STC attention is an **economic** and **effective** way to decompose full spatio-temporal attention.

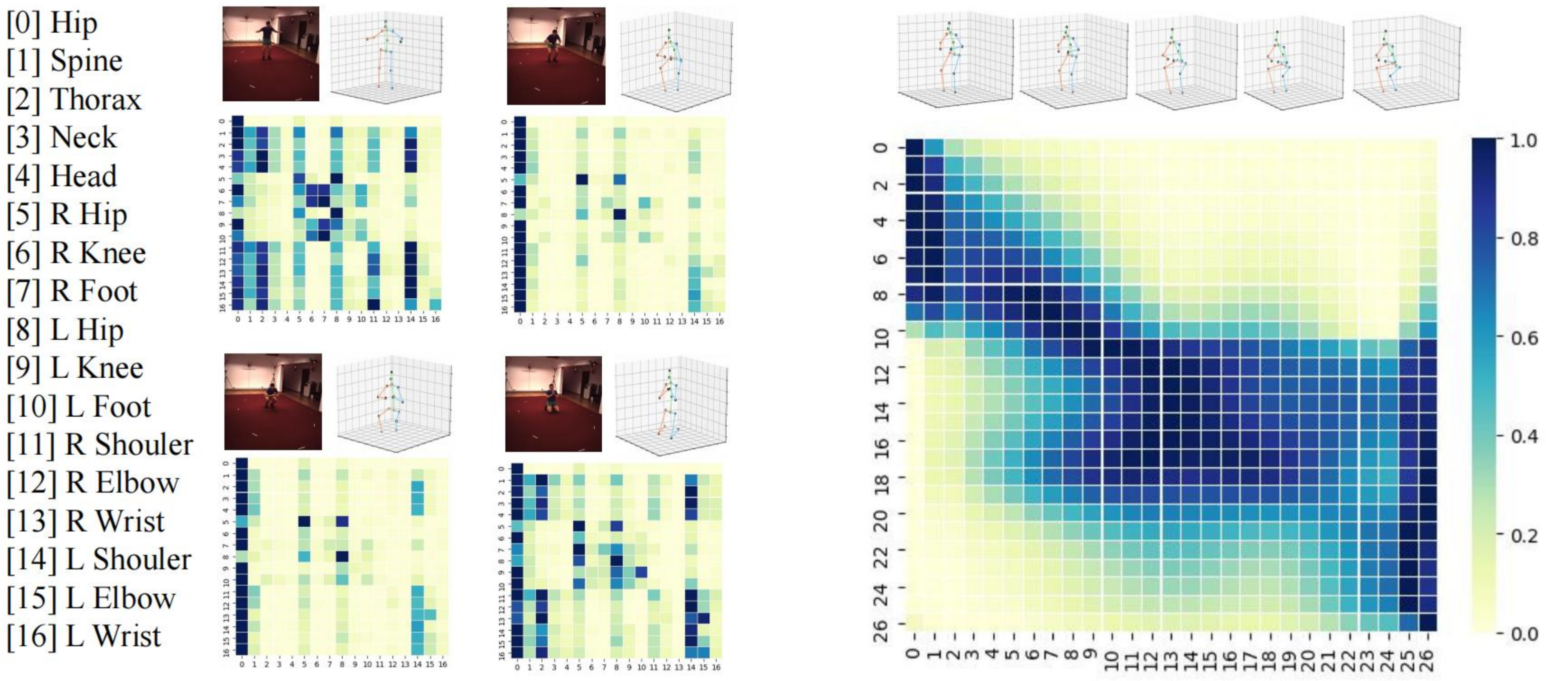
Performance Comparison on MPI-INF-3DHP

Table 3. Performance comparisons in terms of PCK, AUC and P1 with the state-of-the-art methods on MPI-INF-3DHP dataset. Here, the higher PCK, the higher AUC and the lower P1 indicate the better regressions. The best result in each column is marked in red. T is the number of sampled frames from each video.

Method	**Publication**	PCK ↑	AUC ↑	P1(mm) ↓
UGCN [46](T=96)	ECCV'20	86.9	62.1	68.1
Anatomy3D [6] (T=81)	TCSVT'21	87.8	53.8	79.1
PoseFormer [54] (T=9)	ICCV'21	88.6	56.4	77.1
Hu *et al.* [16] (T=96)	ACM MM'21	97.9	69.5	42.5
CrossFormer [13] (T=9)	arXiv'22	89.1	57.5	76.3
PATA [48] (T=243)	TIP'22	90.3	57.8	69.4
MHFormer [23] (T=9)	CVPR'22	93.8	63.3	58.0
MixSTE [52] (T=27)	CVPR'22	94.4	66.5	54.9
Einfalt *et al.* [9] (T=81)	arXiv'22	95.4	67.6	46.9
P-STMO [39] (T=81)	ECCV'22	97.9	75.8	32.2
STCFormer (T=9)		98.2	81.5	28.2
STCFormer (T=27)		98.4	83.4	24.2
STCFormer (T=81)		98.7	83.9	23.1

Our STCFormer with T=81 reaches the to-date best reported performance with PCK of **98.7%**, AUC of **83.9%** and P1 error of **23.1mm**, outperforming the current state-of-the-art models with a large margin of 0.8% in PCK, 8.1% in AUC and 9.1mm in P1 error.

Attention visualization

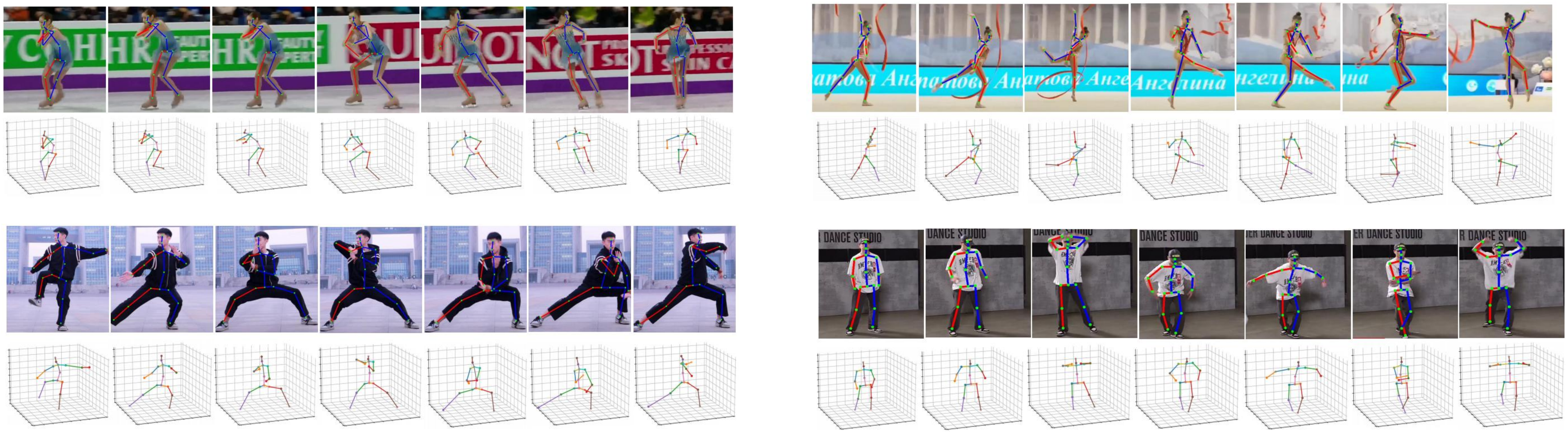


(a) Space

(b) Time

- The spatial attention map shows that our model learns different patterns between joints from the videos of different actions.
- Moreover, the temporal attention map illustrates strong correlation across adjacent frames owing to the continuity of human actions.

Result visualization



Our STCFormer shows great **generalization** ability on in-the-wild videos.

Thank You!
Enjoy your CVPR!

Zhenhua Tang （Presenter）, Zhaofan Qiu, Yanbin Hao, Richang Hong, Ting Yao

Hefei University of Technology
HiDream.ai Inc
University of Science and Technology of China