






An Efficient Spatio-Temporal Pyramid Transformer for Action Detection (Supplementary Materials)

Yuetian Weng¹ , Zizheng Pan¹ , Mingfei Han^{1,2} ,
Xiaojun Chang^{2,3} , and Bohan Zhuang¹ 

¹ Data Science & AI, Monash University

² ReLER Lab, AAIL, University of Technology Sydney

³ School of Computing Technologies, RMIT University

{yuetian.weng, zizheng.pan, bohan.zhuang}@monash.edu, hmf282@gmail.com,
xiaojun.chang@uts.edu.au

We organize our supplementary material as follows.

- In Sec. **A**, we present an illustrated example of LSTA.
- In Sec. **B**, we compare the per-category performance of different models on THUMOS14 and provide attention visualization from the last layer on the two selected action instances.
- In Sec. **C**, we present additional experiment results on ActivityNet 1.3.
- In Sec. **D**, we provide supplementary comparison results with other representative video Transformers.

A Additional Illustration of LSTA

We further illustrate LSTA in Fig. **A**. From left to right, for each input tensor of size $T \times H \times W$, we first evenly divide it into $w_1 \times w_2 \times w_3$ sub-windows, where T , H and W refer to the temporal size and height, width, respectively. Next, each query token only attends to tokens within the same 3D local window. Last, we further reduce both the spatial and temporal resolution of the keys and values within each 3D local window for better efficiency.

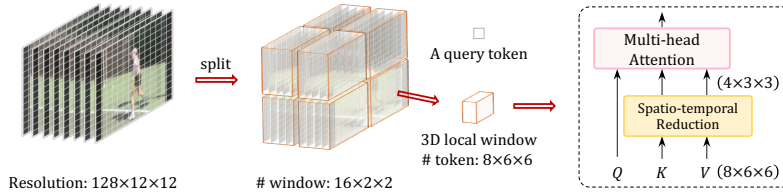


Fig. A: An illustrated example of LSTA.

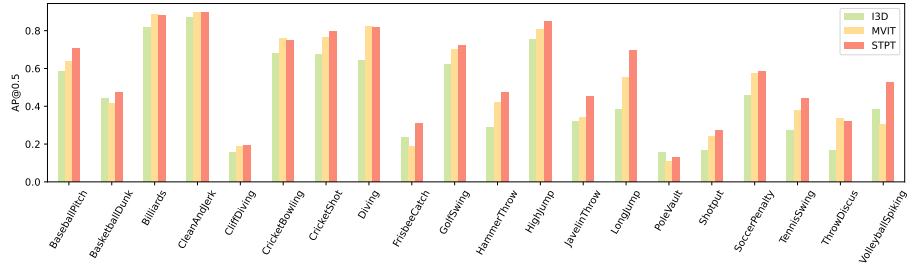


Fig. B: Per-category AP@0.5 on THUMOS14.

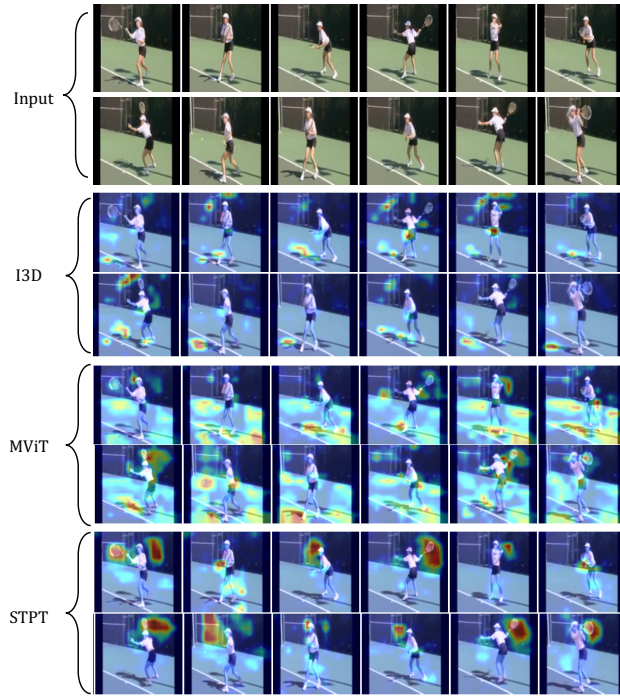


Fig. C: Attention visualization of different models for an action instance of “TennisSwing”.

B Attention Visualization on THUMOS14

In Fig. B, we show the per-category AP@0.5 of I3D, MViT and the proposed STPT on THUMOS14 with RGB input only. It demonstrates that our STPT surpasses the other two models in most categories. It is also notable that our STPT outperforms other models by a large margin on some action categories, *e.g.*, TennisSwing, LongJump.

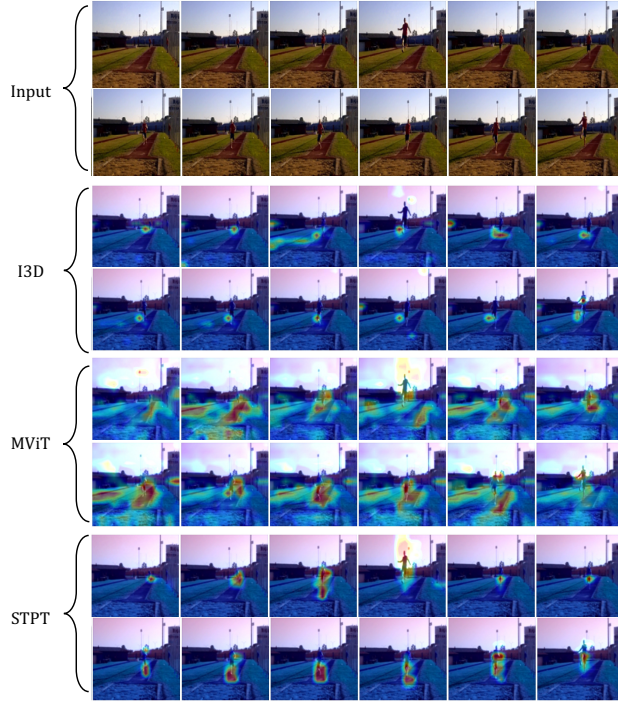


Fig.D: Attention visualization of different models for an action instance of “LongJump”.

We further provide attention visualization of different models on the two example action instances selected from THUMOS14 in Fig. C and Fig. D. For each video instance, we show the sampled input RGB frames and use Grad-CAM to generate the corresponding attention in the last layer. Due to the local receptive field of 3D convolutions, I3D lacks the capacity of learning long-term dependencies in the videos, leading to inaccurate and irrelevant attention. Alternatively, MViT applies global attention in all the stages, which brings noises to the informative spatio-temporal representations and struggles to focus on key objects or actions, *e.g.*, the tennis racket or the jumping action. Different from both cases, by flexibly involving locality constraint in early stages and applying global attention in later stages, our STPT efficiently encodes local patterns and captures global dependencies in a concise manner, enabling learning strong spatio-temporal representations from videos.

C More Results on ActivityNet 1.3

We provide more results on ActivityNet 1.3 in Table A and Table B. In Table A, we report the performance and the computational cost of all the possible

Table A: Effect of our architecture design principle. We evaluate the performance (in mAP) and computational cost (in GFLOPs) of several combinations of blocks on ActivityNet 1.3. L/G refers to LSTA/GSTA used in each stage. Models are equipped with CPE. The number of temporal tokens is set to 96.

Type	GFLOPs	0.5	0.75	0.95	Avg.
LLLL	121.0	49.9	32.4	3.3	31.7
LLLG	126.2	50.0	32.5	3.5	31.8
LGGG	152.2	50.3	32.5	6.8	32.7
GGGG	172.4	50.1	32.7	5.9	32.2
LLGG	134.1	51.4	33.7	6.8	33.4

Table B: Effect of window size in terms of temporal dimension. We compare the performance (in mAP) and computational cost (in GFLOPs) for different scales of a local window in each LSTA block.

Window Size	GFLOPs	0.5	0.75	0.95	Avg.
[1,1,1]	133.9	50.3	33.0	4.0	32.1
[4,4,4]	134.0	50.4	33.2	4.1	32.3
[8,8,8]	134.1	51.0	33.5	5.7	32.9
[8,8,16]	134.1	51.4	33.7	6.8	33.4
[16,16,16]	134.4	49.7	32.0	5.6	32.0

Table C: More comparisons (mAP(%) at different tIoU thresholds) with other ViT models on THUMOS14 and ActivityNet 1.3.

Backbone	GFLOPs	0.3	0.4	0.5	0.6	0.7	Avg.
DualFormer	112.2	68.5	63.0	54.4	41.8	27.7	51.1
RegionViT	181.3	68.6	62.7	53.5	41.7	28.8	51.1
Twins	119.8	69.7	62.7	54.1	43.2	28.7	51.7
Uniformer	134.2	68.7	63.6	54.6	42.0	28.7	51.5
Ours	111.2	70.6	65.7	56.4	44.6	30.5	53.6

(3) THUMOS14

Backbone	GFLOPs	0.5	0.75	0.95	Avg.
DualFormer	171.1	50.7	33.1	5.4	32.7
RegionViT	241.1	51.3	33.4	6.2	32.8
Twins	140.2	51.1	33.2	5.2	32.7
Uniformer	185.5	50.7	32.8	5.2	32.6
Ours	134.1	51.4	33.7	6.8	33.4

(3) ActivityNet 1.3

combinations of LSTA (L) and GSTA (G) on ActivityNet 1.3. Without using GSTA, the model LLLL is computationally efficient but lacks the capacity of learning global dependency, resulting in the lowest mAP scores compared with other structures. Alternatively, applying GSTA in all the stages leads to heavy computational cost (134.1G *vs.* 172.4G). However, the mAP scores drop at all thresholds as the model cannot extract detailed spatio-temporal patterns in the early stages. Thus, for ActivityNet, we choose LSTA and GSTA in the first two stages and the last two stages respectively, in order to achieve a favourable balance between efficiency and effectiveness. As shown in Table B, with comparable FLOPs, the model with the window size of [8,8,16] for LSTA outperforms the other settings at all thresholds.

D Comparison with other video Transformers

Compared with DualFormer, RegionViT and Twins, which alternatively process local and global information within each block, we leverage local LSTA in the early stages to remove local redundancy and utilize global GSTA in the deeper layers to model the long-term dependencies. Different from the factorized space-time attention used in RegionViT, we encode the target motions by jointly aggregating spatio-temporal relations. As shown in Table C, our STPT consistently achieves better performance with fewer FLOPs than other methods on both datasets.