Paper ID 659

SMU SINGAPORE MANAGEMENT UNIVERSITY

之江实验室 ZHEJIANG LAB

ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

Long-term Leap Attention, Short-term Periodic Shift for Video Classification

Hao Zhang, Lechao Cheng, Yanbin Hao*, Chong-Wah Ngo

• Motivation

➢ Video transformer processes T times longer sequence than the vision transformer.

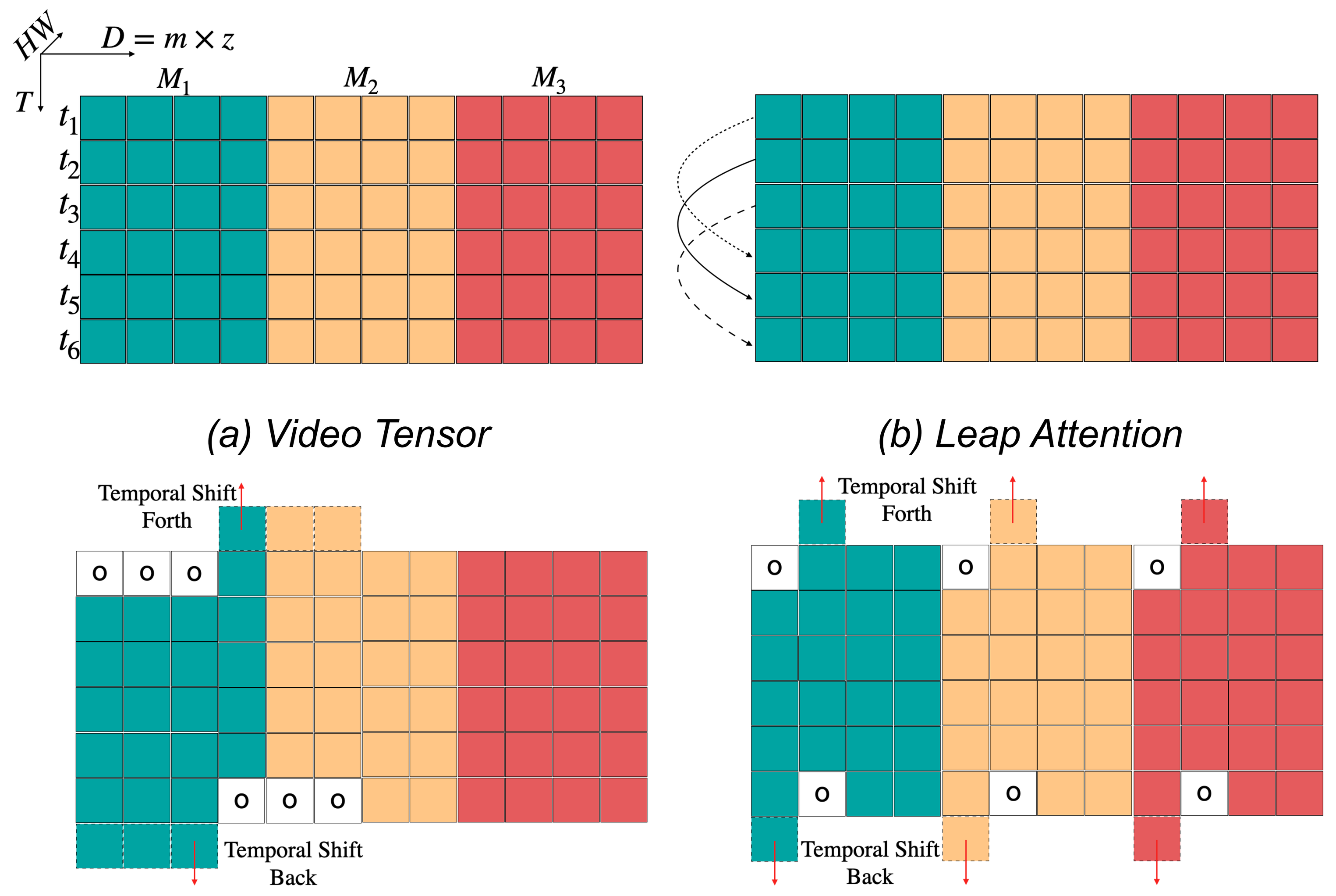
	Tensor Shape	Distance Calculations
Image	$z_i \in \mathbb{R}^{N \times D}$	N^2
Video	$z_v \in \mathbb{R}^{T \times N \times D}$	T^2N^2

➢ Temporally ***neighboring frames*** are generally ***similar*** (redundancy) despite being different in ***micro details***.

➢ To avoid redundancy, we can suppress attention on visually similar frames in a dilated manner. For micro details, we can launch short-term temporal shift operation. Complexity becomes: $T^2N^2 \rightarrow 2TN^2$

• Proposed Framework

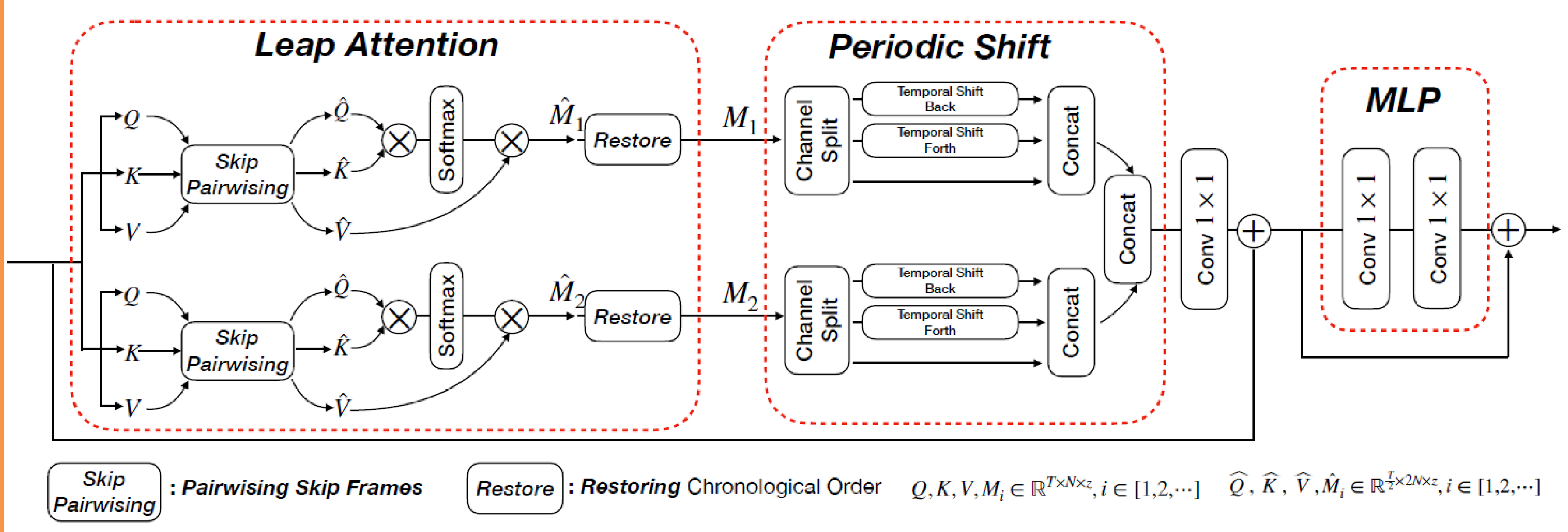
➢ ***Leap Attention with Periodic Shift encoder (LAPS)***: contains Leap Attention (LA) and Periodic Shift (PS). The LA and PS separately serves to model long-term temporal relations and short-term variations between adjacent frames.



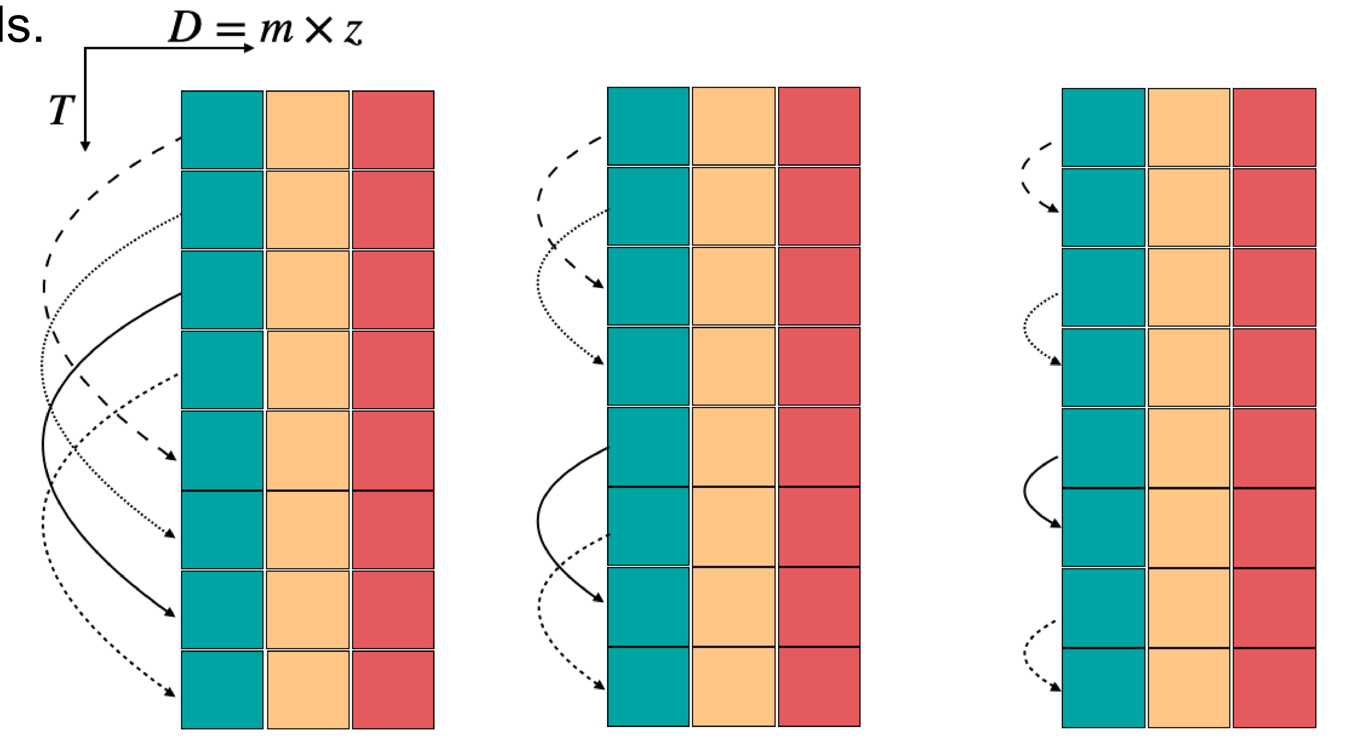
(a) Video Tensor *(b) Leap Attention*

(c) Plain Shift *(d) Periodic Shift*

➢ ***An LAPS overview***: is a zero-parameter, lightweight-FLOPs attention alternative. It can flexibly replace a generic 2D attention and convert a static vision transformer into a video one.



➢ ***Pyramid Skipping*** aims to connect frames with various distances into pairs, therefore facilitating the LA to have multi-scale temporal receptive fields. $D = m \times z$



(a) Pyramid-1. *(b) Pyramid-2* *(c) Pyramid-3*

• Experiments

➢ ***Ablation Study***:

1. LAPS *vs* 2/3D Attention.

Model	MSHA	GFLOPs	Params (M)	Top-1 (%)
Base2D	2D Atten	39.1	39.8	74.00
Base3D	3D Atten	46.5 (18.9% ↑)	39.8	76.31
LAPS	Plain Shift	39.1	39.8	74.86
	P-Shift	39.1	39.8	75.19
	LA	40.1 (2.6% ↑)	39.8	75.84
	P-Shift + LA	40.1 (2.6% ↑)	39.8	76.04

2. Pyramid Skipping

Model	Pyramids	Skipped Steps	Top-1
LAPS	R=[3, 3, 3]	S=[1/8, 1/8, 1/8] · T	75.55
	R=[2, 2, 2]	S=[1/4, 1/4, 1/4] · T	75.82
	R=[1, 1, 1]	S=[1/2, 1/2, 1/2] · T	75.86
	R=[1, 2, 3]	S=[1/2, 1/4, 1/8] · T	**76.04**

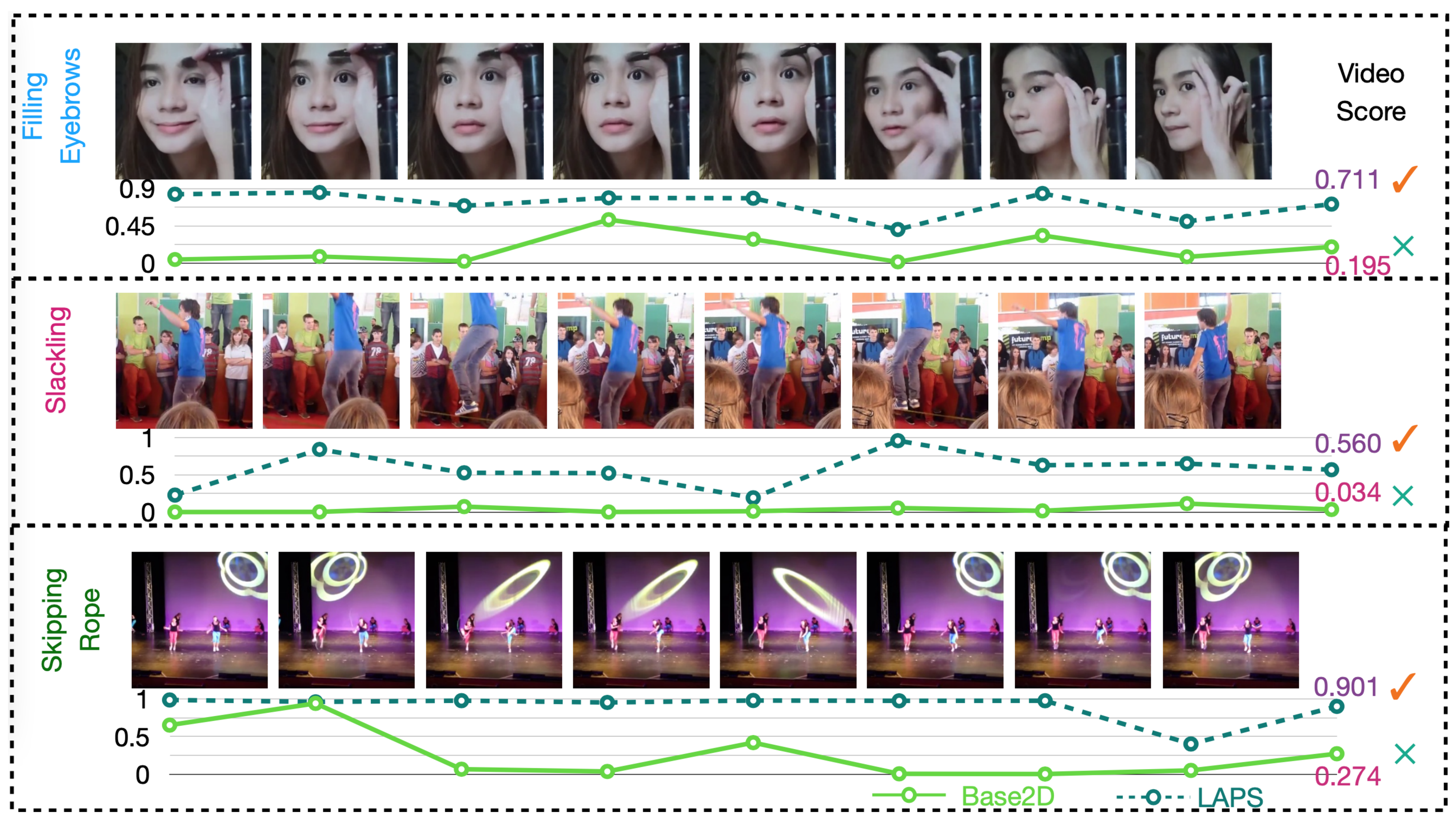
➢ ***Comparison with SOTA***:

Model	Base	Pretrain	#F×Res ($T \times HW$)	GFLOPs×Views	Params (M)	Training Epochs	Top-1 (%)	Top-5 (%)
TDN-R50 [39]	ResNet50	IN-1K	24×256^2	108.0×30	26.6	100	78.40	93.60
TDN-R101 [39]	ResNet101	IN-1K	24×256^2	198.0×30	43.9	100	79.40	94.40
GC-TDN-R50 [14]	ResNet50	IN-1K	24×256^2	110.1×30	27.4	100	79.60	94.10
SlowFast 8 × 8 [9]	ResNet50	None	32×256^2	65.7×30	-	196	77.00	92.60
SlowFast 16 × 8 [9]	ResNet101+NL	None	32×256^2	234.0×30	59.9	196	79.80	93.90
X3D-L [8]	X2D	None	16×356^2	24.8×30	6.1	256	77.50	92.90
X3D-XL [8]	X2D	None	16×356^2	48.4×30	11.0	256	79.10	93.90
ViT (Video) [47]	ViT-B	IN-22K	8×224^2	134.7×30	85.9	18	76.00	92.50
TokShift [47]	ViT-B	IN-22K	16×224^2	269.5×30	85.9	18	78.20	93.80
TokShift (MR) [47]	ViT-B	IN-22K	8×256^2	175.8×30	85.9	18	77.68	93.55
VTN [28]	ViT-B	IN-22K	250×224^2	4218.0×1	114.0	25	78.60	93.70
TimeSformer [1]	ViT-B	IN-22K	8×224^2	590.0×3	121.4	15	78.00	93.70
Video Swin [25]	Swin-B	IN-1K	32×224^2	281.6×12	88.0	30	80.60	94.60
MViT [6]	MViT-B	None	64×224^2	455.0×9	36.64	200	81.20	95.10
LAPS	Visformer	IN-10K	8×224^2	40.1×15	39.8	18	76.04	92.56
LAPS (L)	Visformer	IN-10K	16×320^2	173.0×15	40.0	18	78.71	93.77
LAPS (H)	Visformer	IN-10K	32×320^2	346.0×15	40.0	18	79.72	94.08
LAPS (E)	Visformer	IN-15K	32×360^2	434.0×15	40.2	18	80.03	94.48

Table 4: Comparison to state-of-the-arts on Kinetics-400 Val.

• Visualization

➢ ***Visualization of Video Exemplars***



We test Base2D and LAPS on the Kinetics-400 val set. The solid and dashed line separately denotes per-frame predictions from Base2D and LAPS transformer.

• Conclusion and Resources

➢ ***Conclusions***:

1. LAPS is a cost-effective alternative to 3D attention for temporal modeling.
2. We could build long/short-term relation with Leap Attention/Periodic Shift.
3. Leap Attention is a new temporal dilated attention.

➢ ***Contact & Resources***

zhanghaoinf@gmail.com
chenglc@zhejianglab.com
haoyanbin@hotmail.com
cwngo@smu.edu.sg

GitHub

PDF Adobe