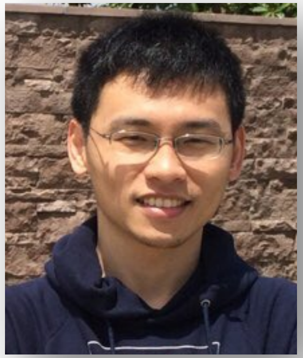
中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES

Baidu 百度

CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

Dynamic Inference: A New Approach Toward Efficient Video Action Recognition





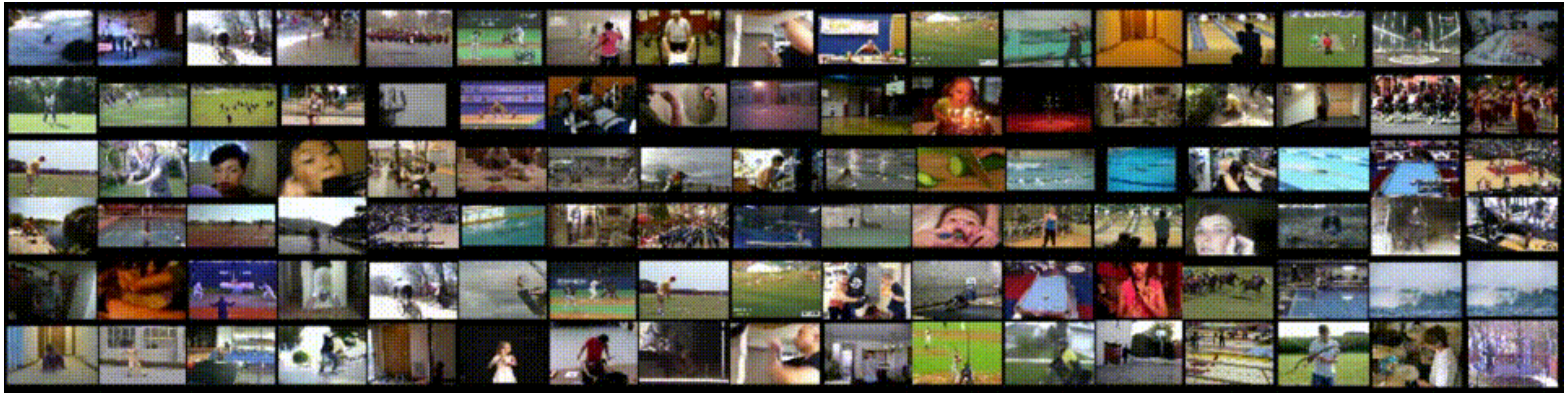

Wenhao Wu *Dongliang He* *Xiao Tan* *Shifeng Chen* *Yi Yang* *Shilei Wen*

Task

Action Recognition: classify the short clip or untrimmed video into pre-defined class.

Task

Action Recognition: classify the short clip or untrimmed video into pre-defined class.



- More than simply recognizing objects
- Complex person-person interaction & people-object interactions
- Videos bring motions

Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

Direction 1: Lightweight Base Model

1. *2D Conv + Efficient Temporal Modeling*
2. *Decompose 3D Conv*
3. *Network architecture search*
4. *Others …*

Motivation

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

Direction 1: Lightweight Base Model	***Direction 2: Adaptive Frame Sampler***
1. 2D Conv + Efficient Temporal Modeling 2. Decompose 3D Conv 3. Network architecture search 4. Others …	• Hand-crafted sampler: Uniform sampling, Dense sampling • Adaptive frame sampler: Adaframe [1], MARL [2], SCSampler [3]

[1] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proc. CVPR*, 2019.
[2] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proc. ICCV*, 2019.
[3] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proc. ICCV*, 2019.

Motivation

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

How to Improve the Computation Efficiency of Action Recognition in Videos?

The time to process one frame AND the number of processed frames.

Direction 1: Lightweight Base Model

1. *2D Conv + Efficient Temporal Modeling*
2. *Decompose 3D Conv*
3. *Network architecture search*
4. *Others …*

Direction 2: Adaptive Frame Sampler

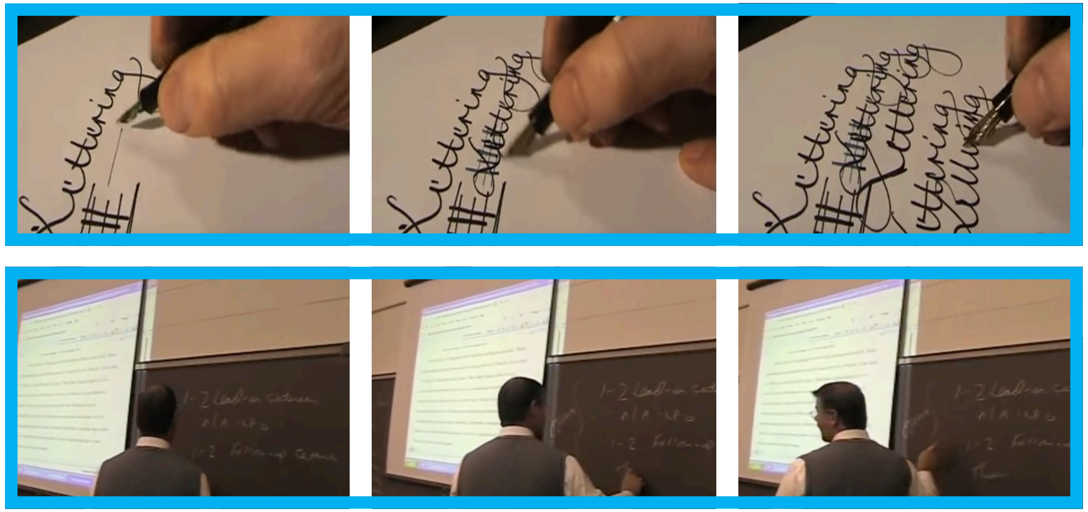
- *Hand-crafted sampler:* *Uniform sampling, Dense sampling*
- *Adaptive frame sampler:* *Adaframe [1], MARL [2], SCSampler [3]*

Direction 3: Dynamic Network Route

- *Image recognition:* *MSDNet [4], SkipNet [5] …*
- *Video recognition:* *We try to improve efficiency from dynamic inference viewpoint.*

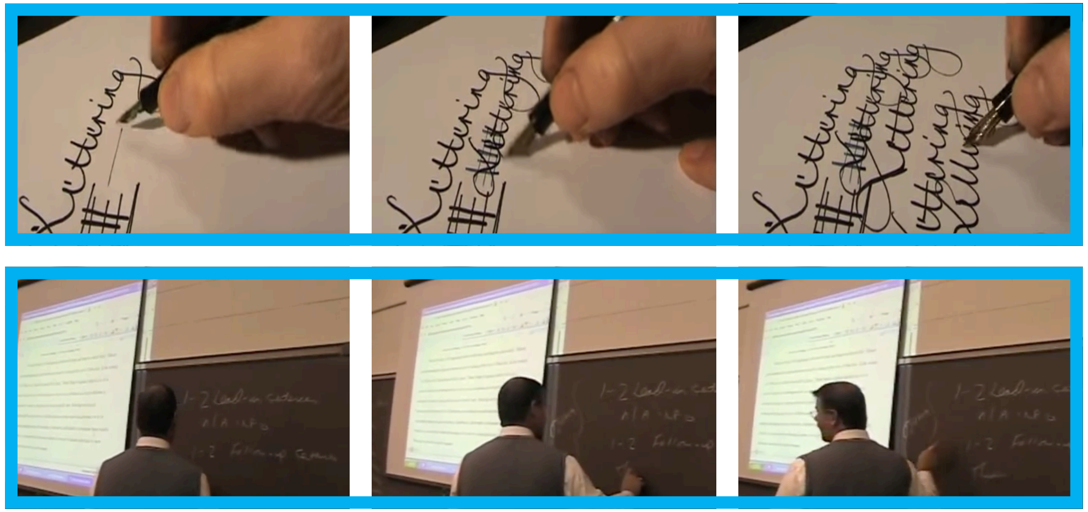
[1] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proc. CVPR*, 2019.
[2] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proc. ICCV*, 2019.
[3] Bruno Korbar, Du Tran, and Lorenzo Torresani. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proc. ICCV*, 2019.
[4] Gao Huang and Danlu Chen. Multi-scale dense networks for resource efficient image classification. In *Proc. ICLR*, 2018.
[5] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proc. ECCV*, 2018.

Observation



(a) Different “Writing” video instances

Observation

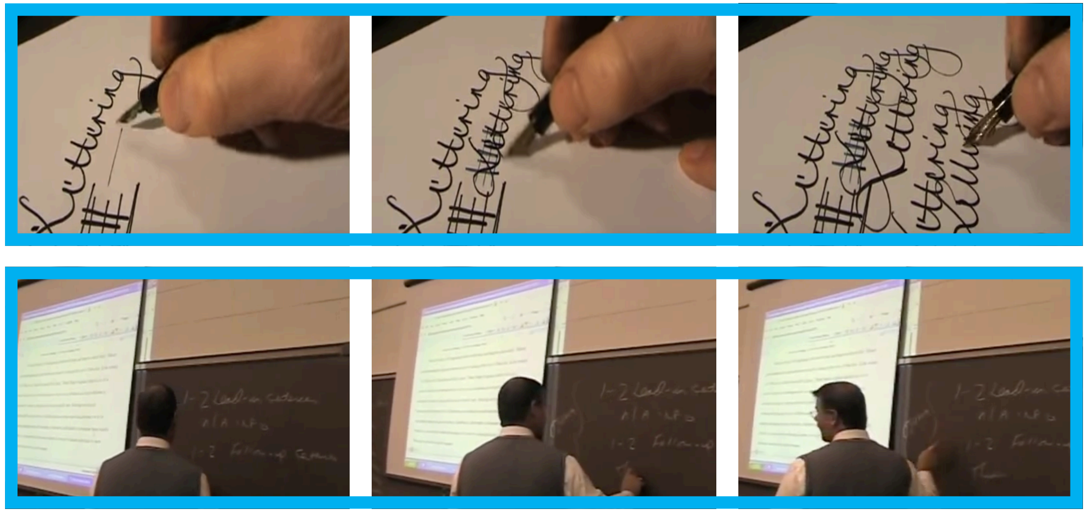


(a) Different "Writing" video instances

Irregular viewpoint

Need varying network capability

Observation



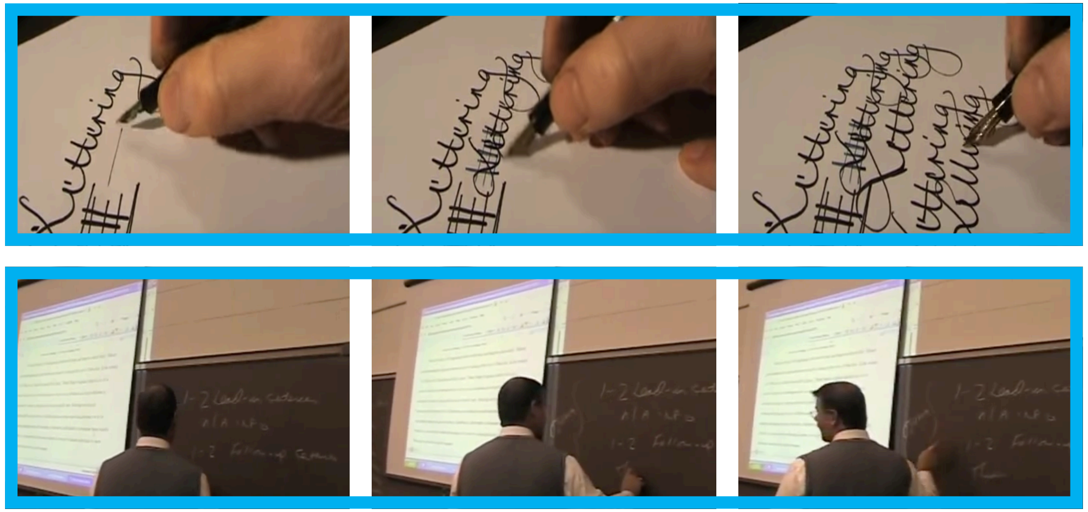
(a) Different "Writing" video instances

(b) "Running" vs. "Long Jump"

Irregular viewpoint

Need varying network capability

Observation



(a) Different “Writing” video instances

Irregular viewpoint

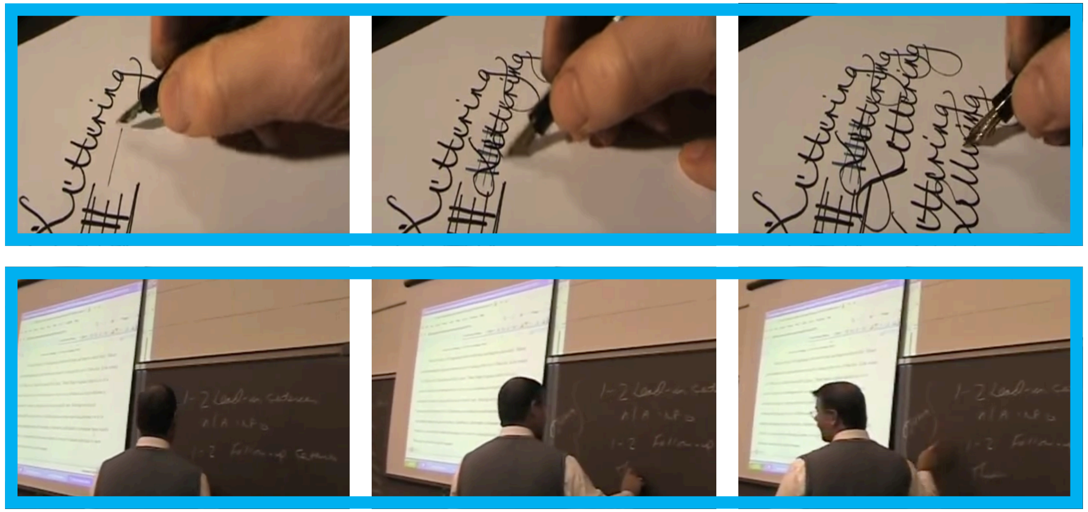
Need varying network capability

(b) “Running” vs. “Long Jump”

Different from “Writing”

Need varying number of frames

Observation



(a) Different "Writing" video instances

(b) "Running" vs. "Long Jump"

Irregular viewpoint

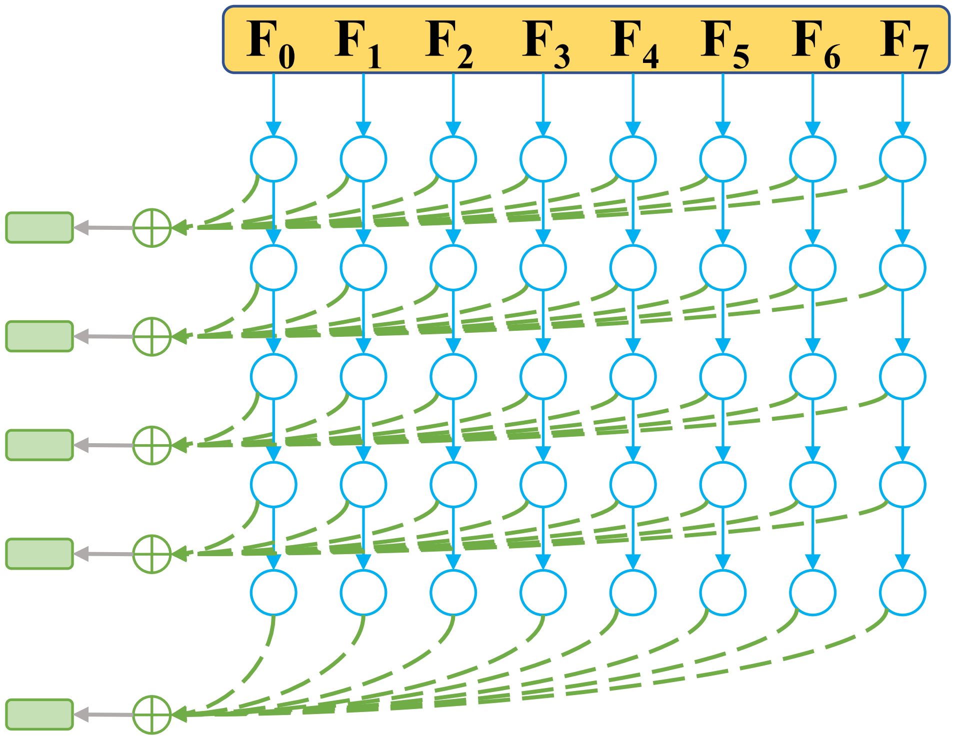
Need varying network capability

Different from "Writing"

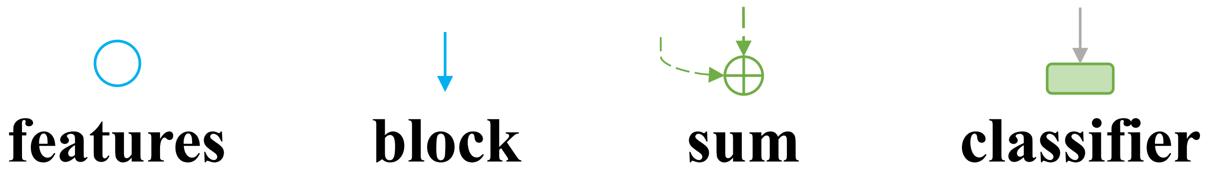
Need varying number of frames

Videos differentiate from each other in terms of their distinguishability.

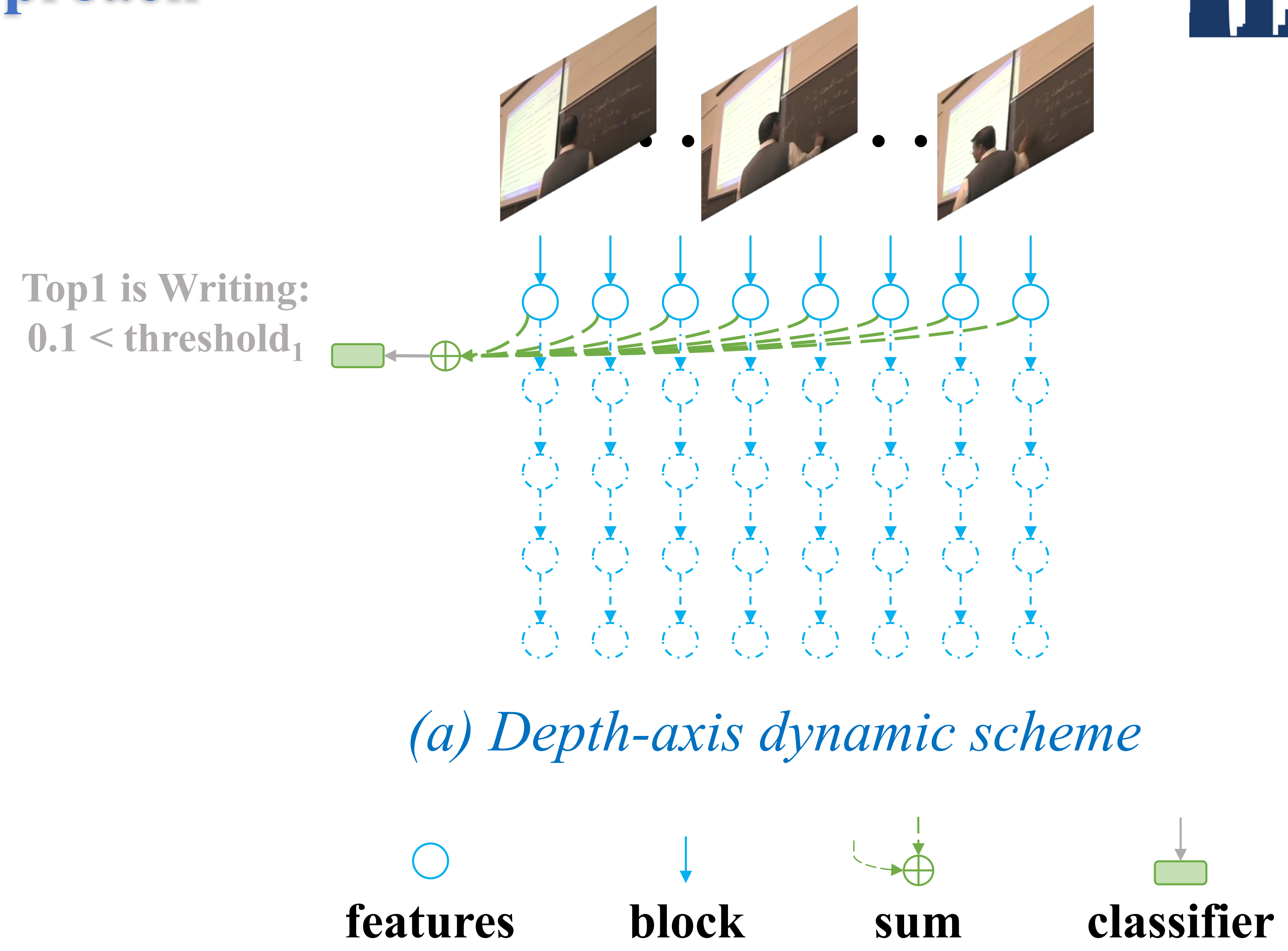
Approach



(a) Depth-axis dynamic scheme

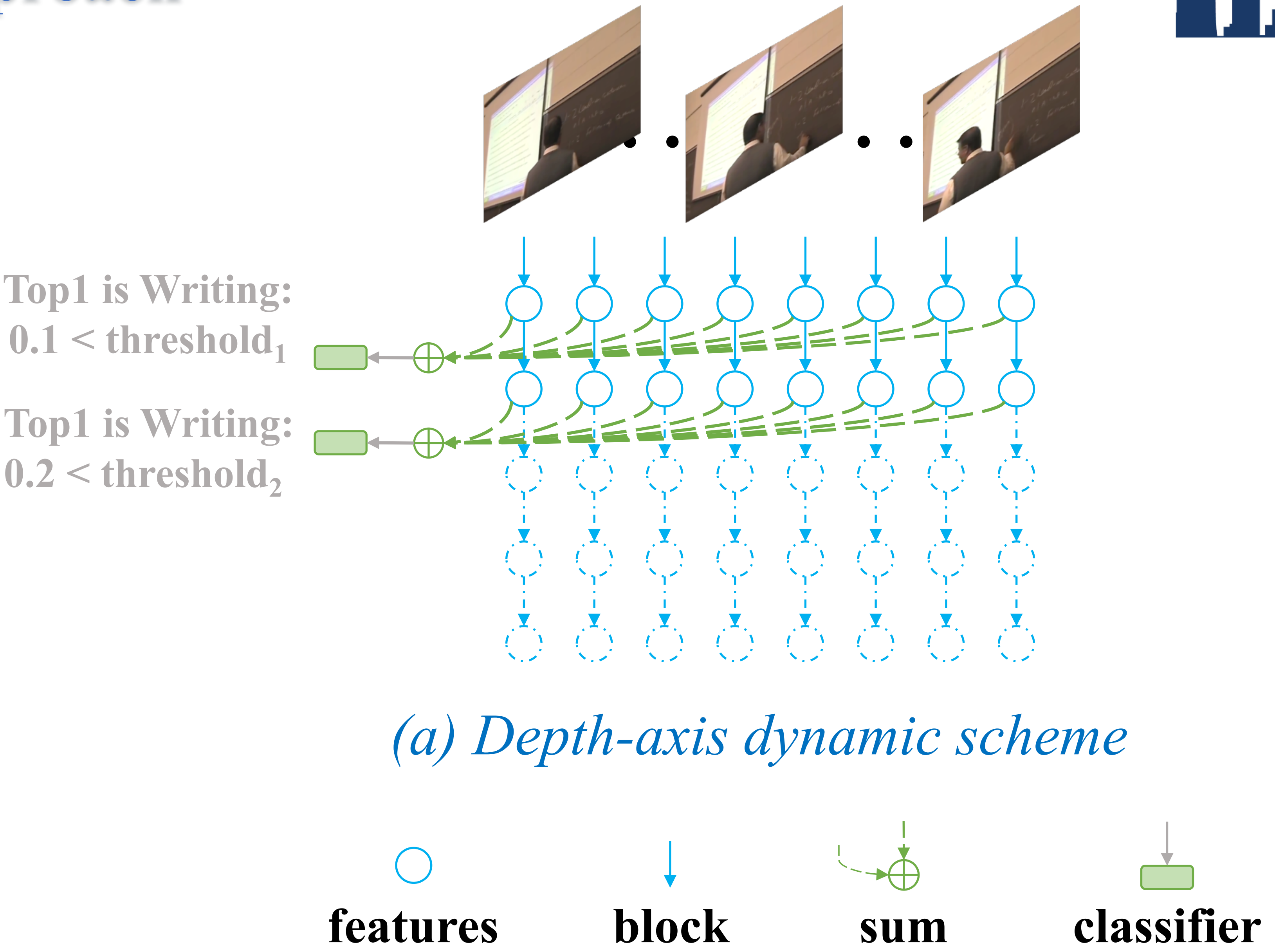


Approach



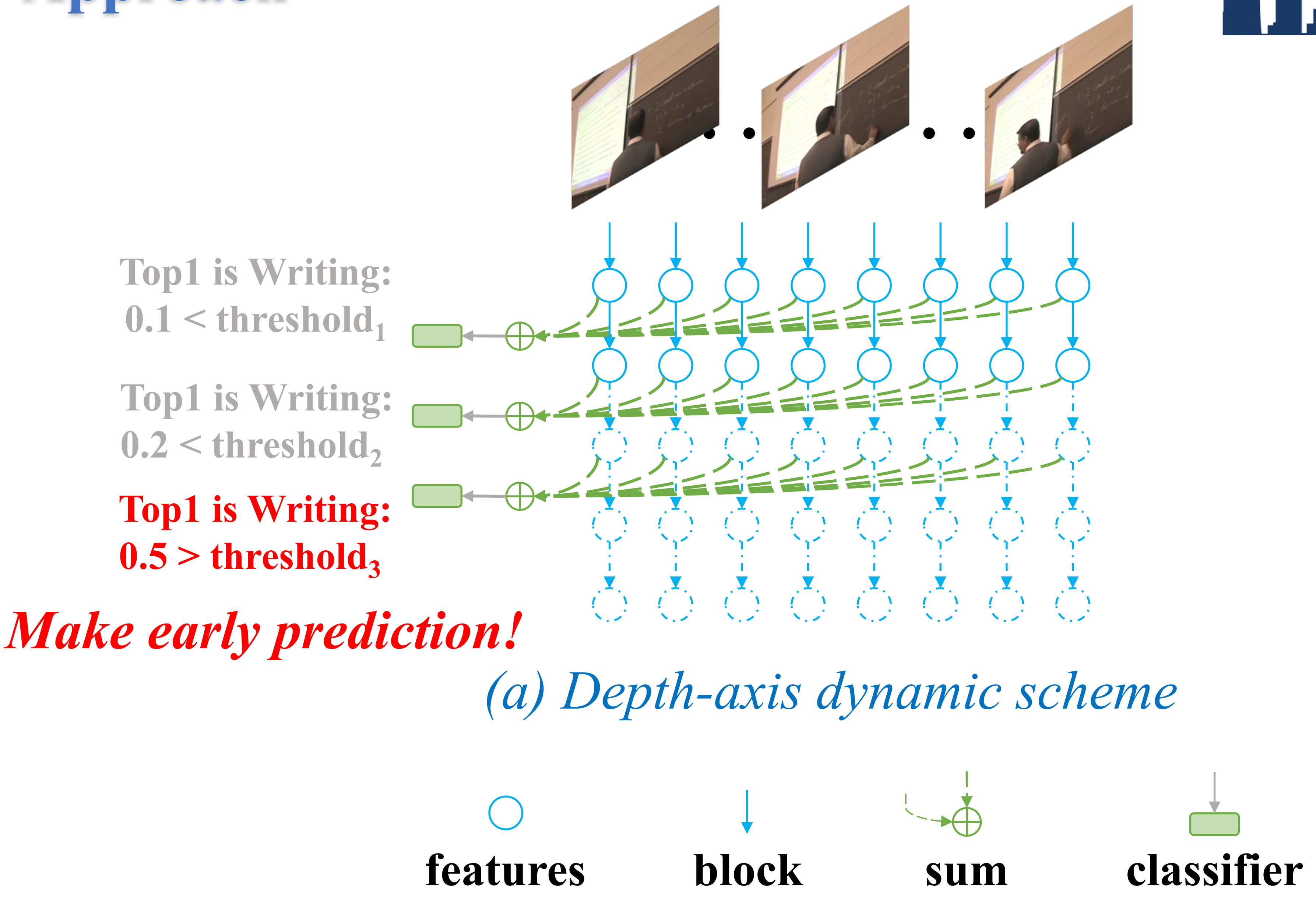
(a) Depth-axis dynamic scheme

Approach



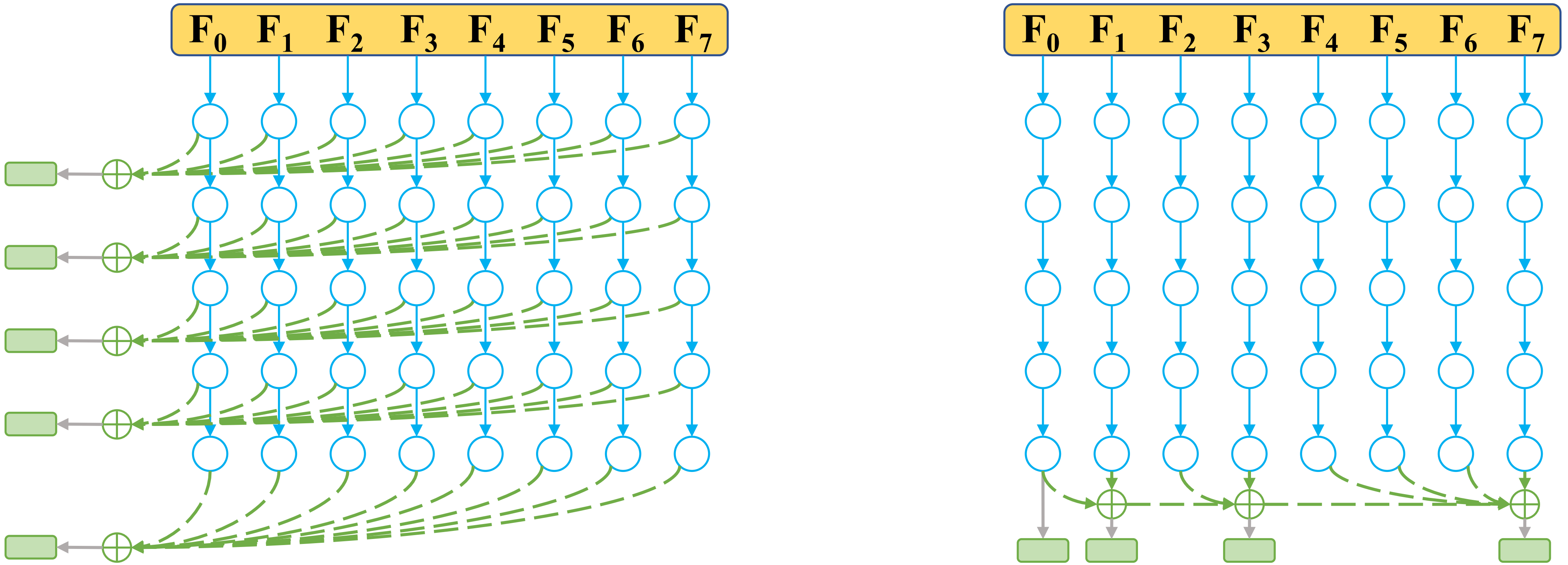
(a) Depth-axis dynamic scheme

Approach



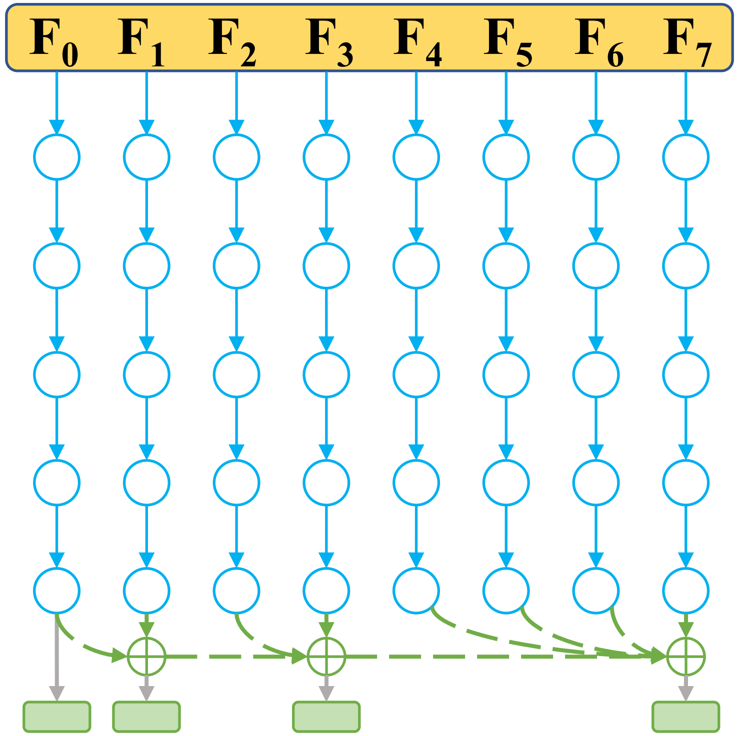
(a) Depth-axis dynamic scheme

Approach

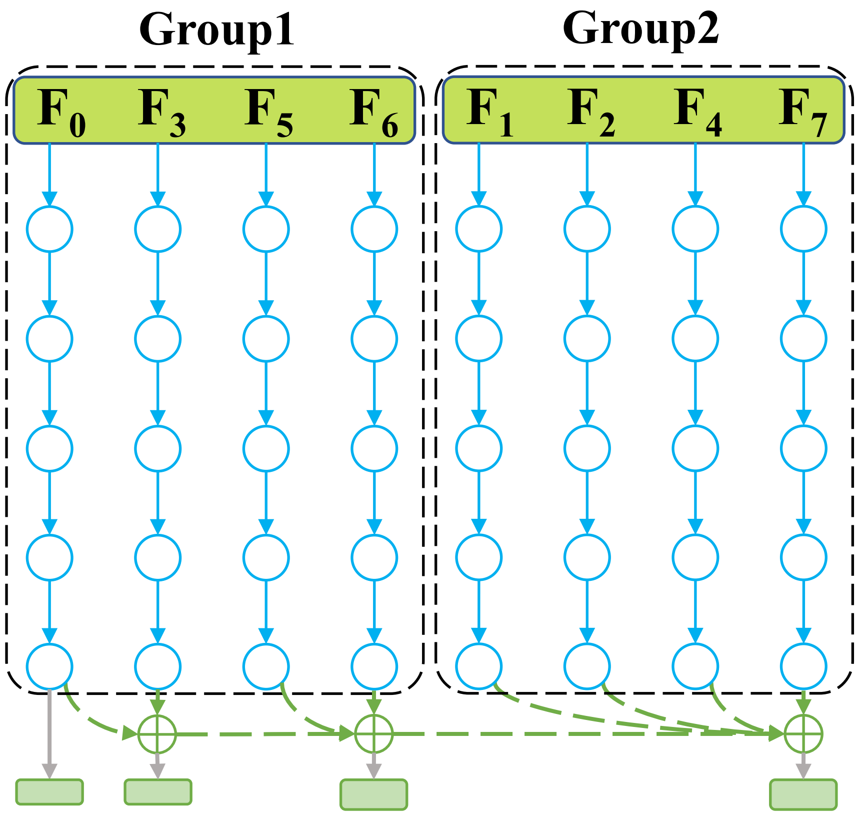


(a) Depth-axis dynamic scheme *(b) Input-axis dynamic scheme*

Approach

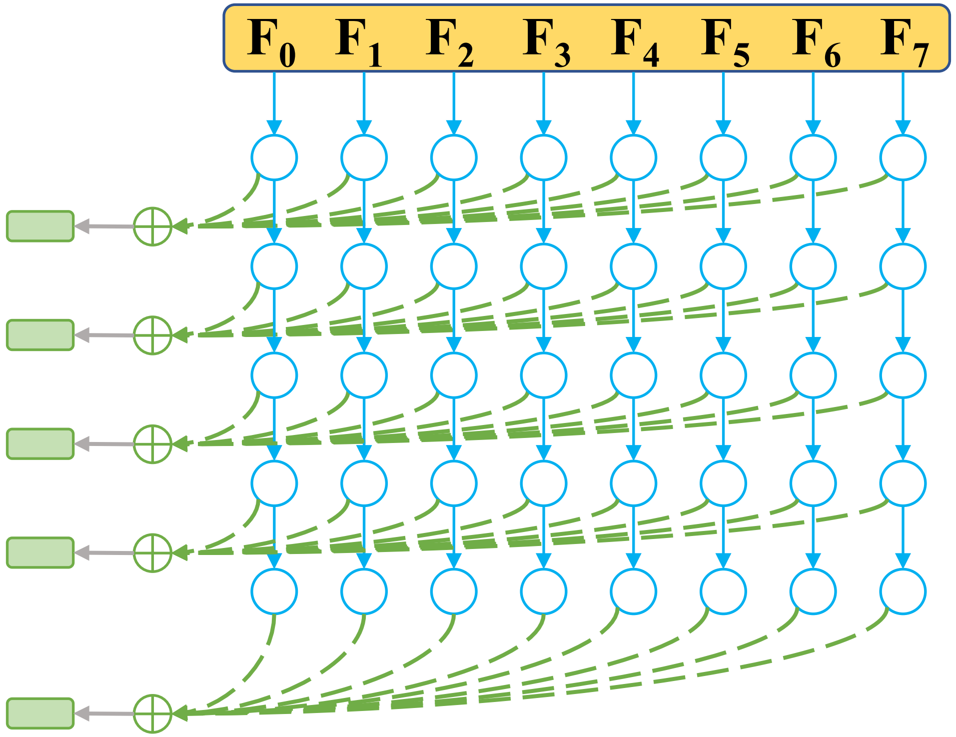


(b) Input-axis dynamic scheme



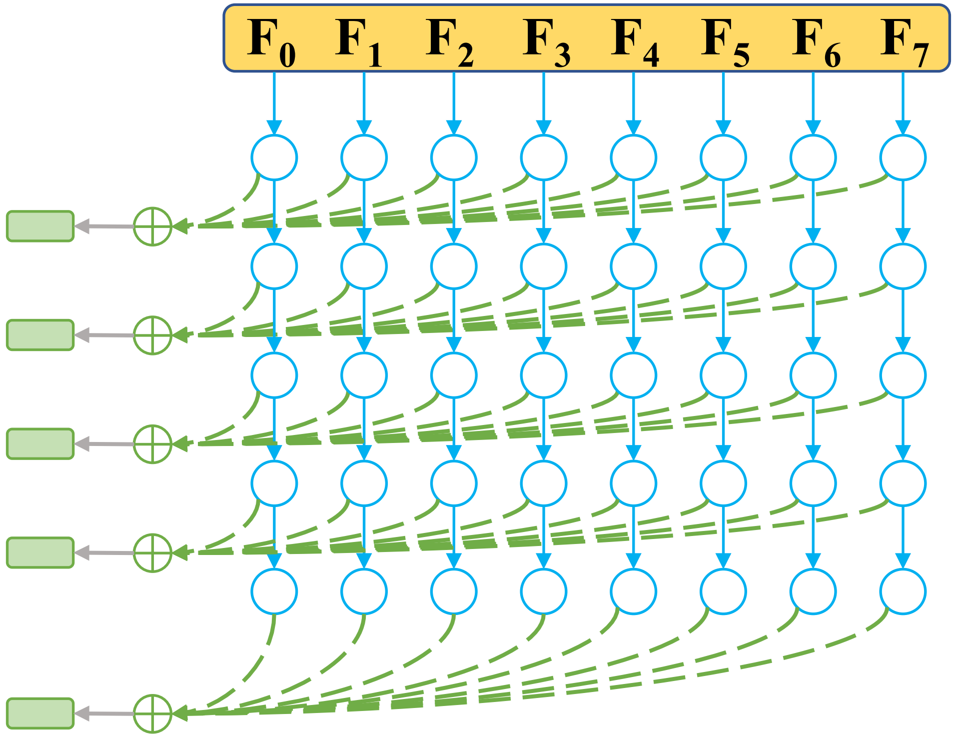
(c) Input with permutation

Approach



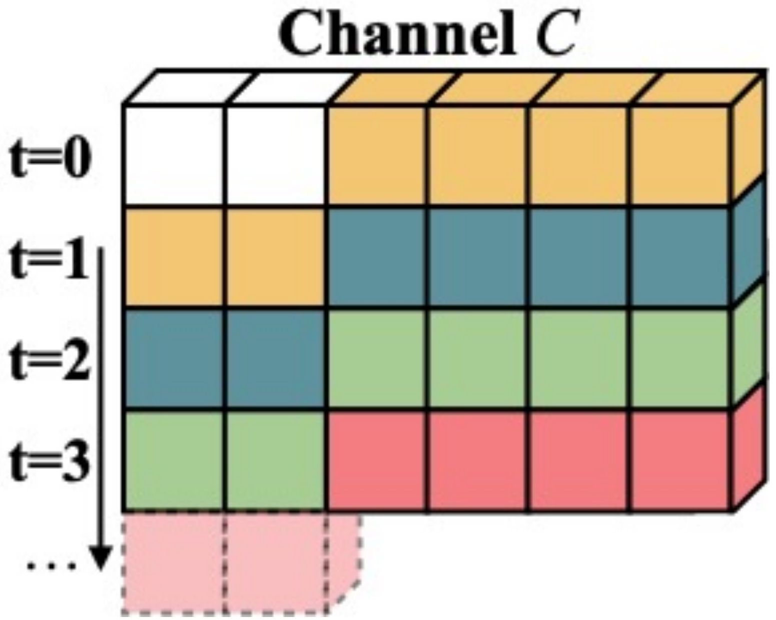
(a) Depth-axis dynamic scheme

Approach

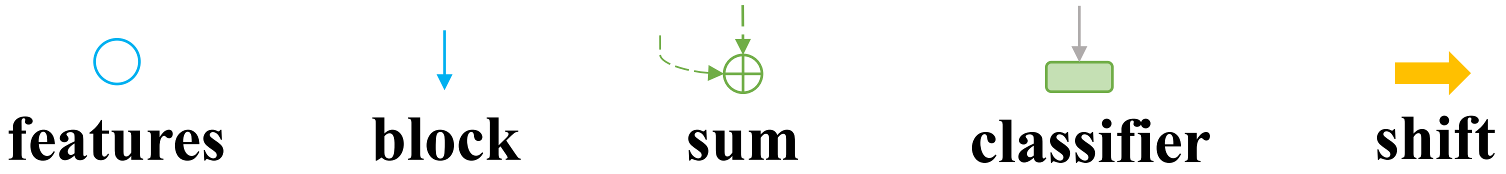


(a) Depth-axis dynamic scheme

Online temporal shift module [1]

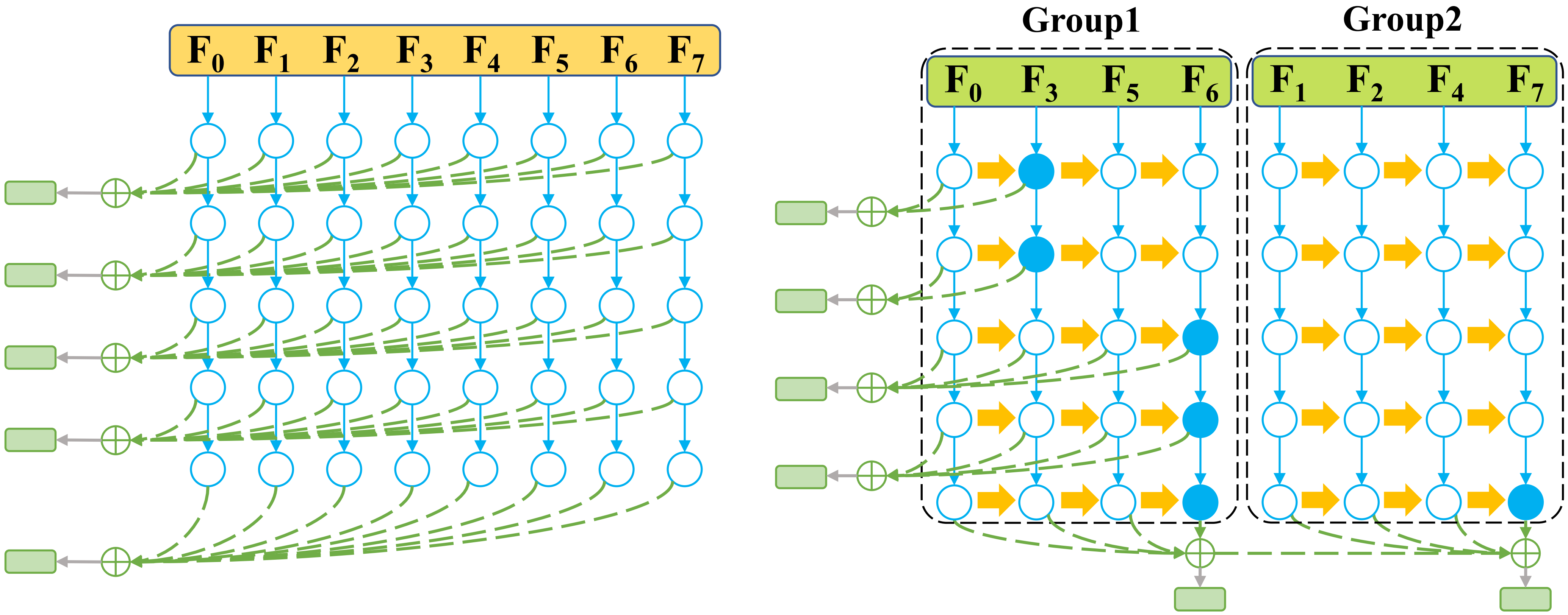


[1] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proc. ICCV*, 2019.



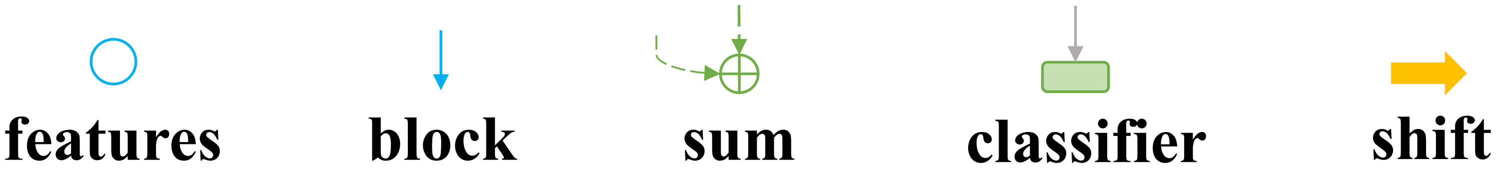
Approach

CVPR SEATTLE WASHINGTON JUNE 16-18 2020



(a) Depth-axis dynamic scheme

(d) Unified



Experiment

Scene-related Datasets

- Kinetics-400
  - 306,245 videos, 400 activity classes
- UCF-101
  - 13,320 videos, 101 classes
- HMDB-51
  - 6,766 videos, 51 classes

Temporal-related Datasets

- Something-Something
  - V1 : 108,499 videos, 174 classes
  - V2 : 220,847 videos, 174 classes

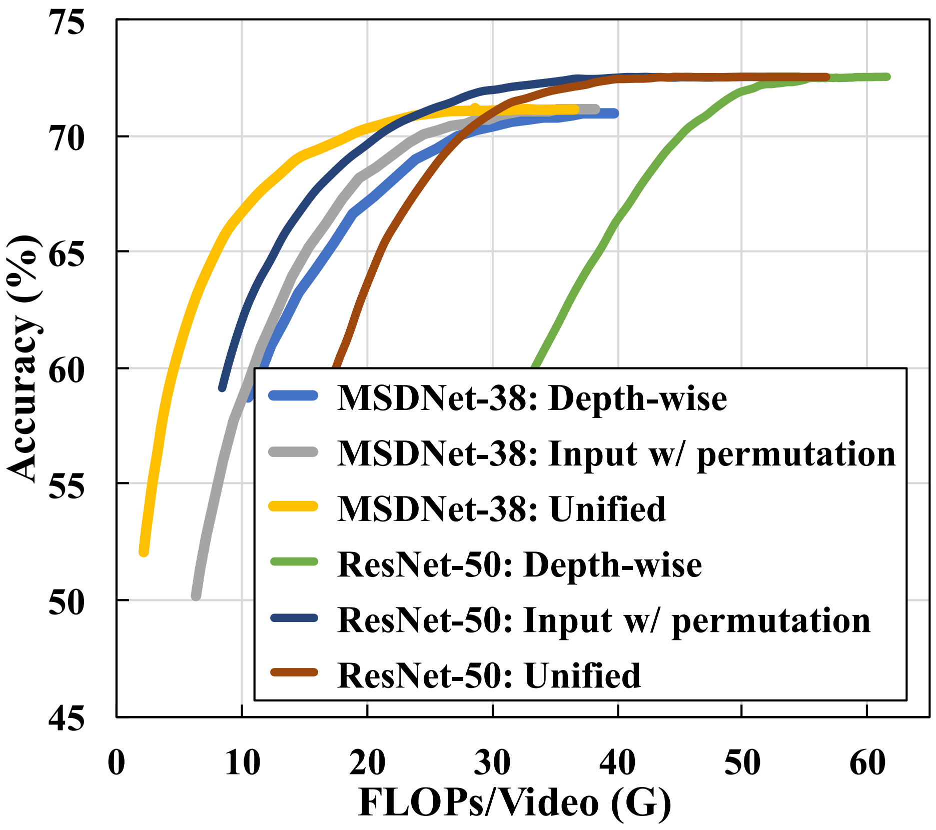
Evaluation Metrics

- Top-1 precision
- Average FLOPs/Video

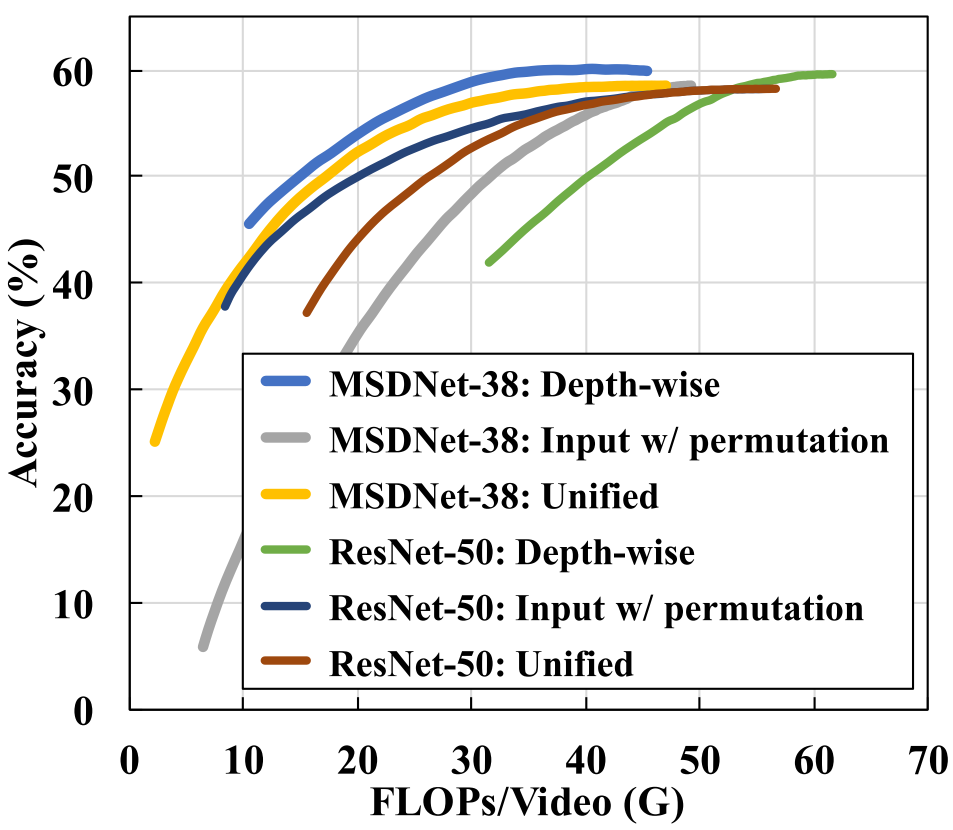
FLOPs, which is short for float-point operations

Experiment

Instantiation: MSDNet-38 [1] & ResNet-50 [2]



(a) Kinetics-400

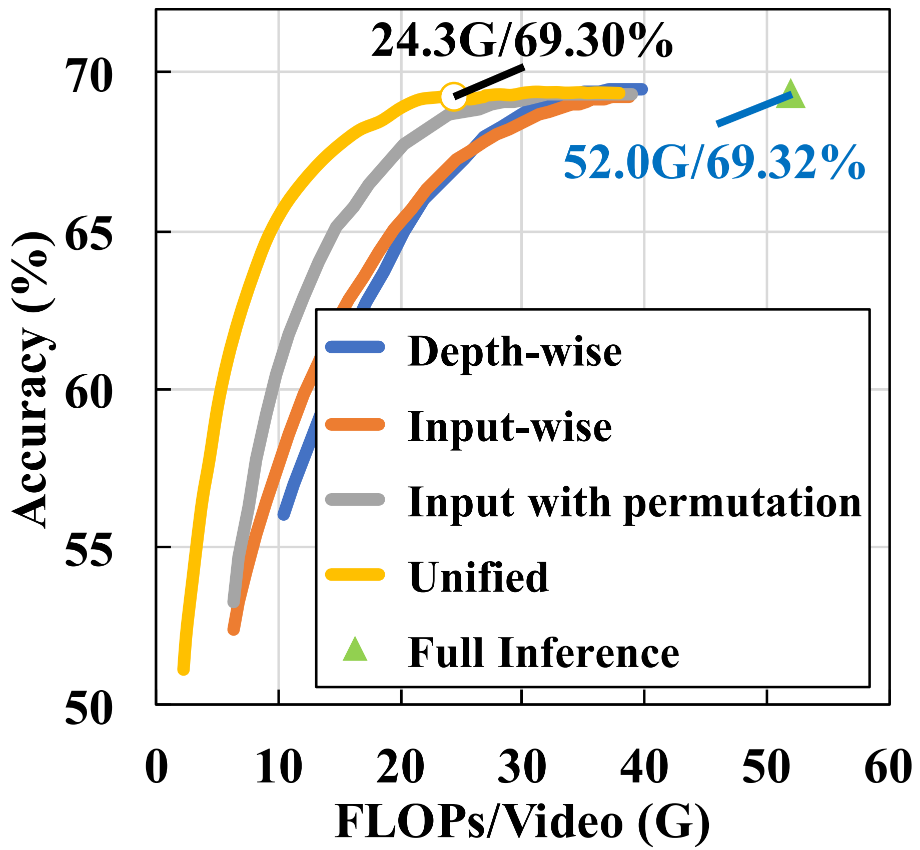


(b) Something-Something v2

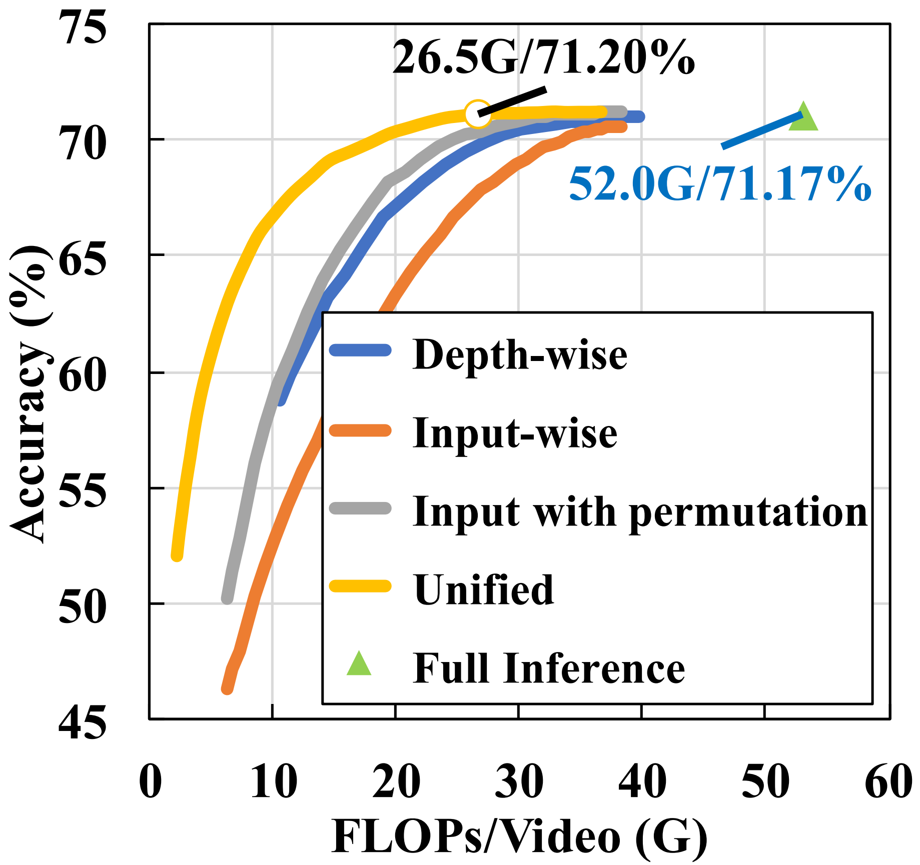
[1] Gao Huang and Danlu Chen. Multi-scale dense networks for resource efficient image classification. In *Proc. ICLR*, 2018

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

Experiment: Ablation Study



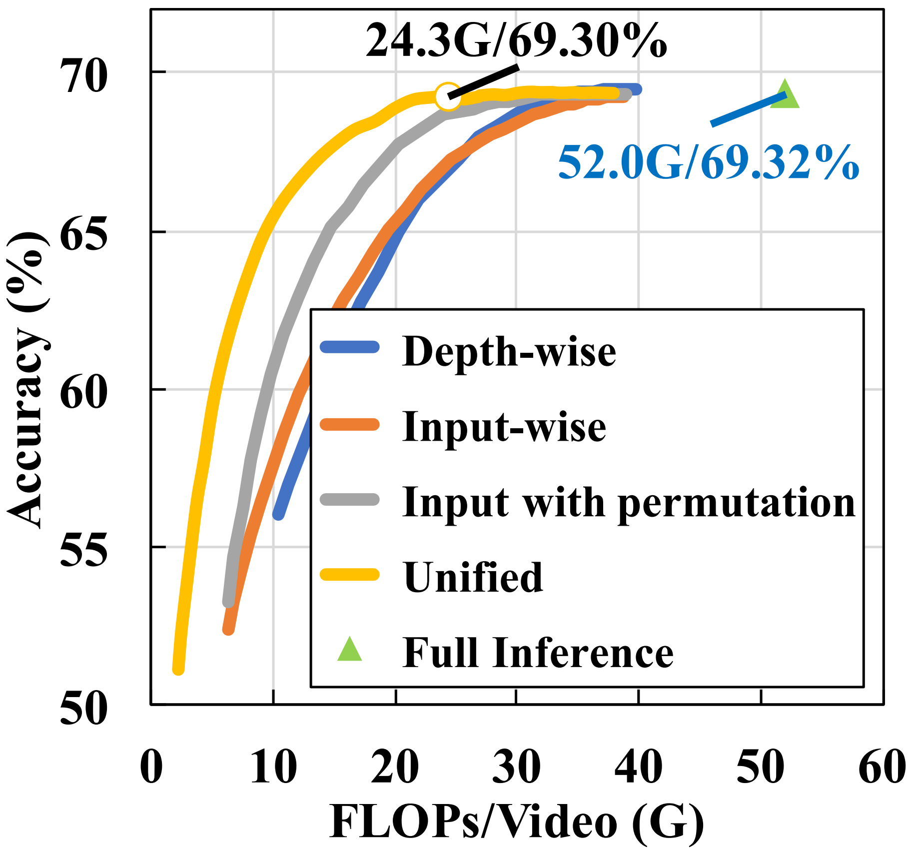
(a) W/o online temporal shift



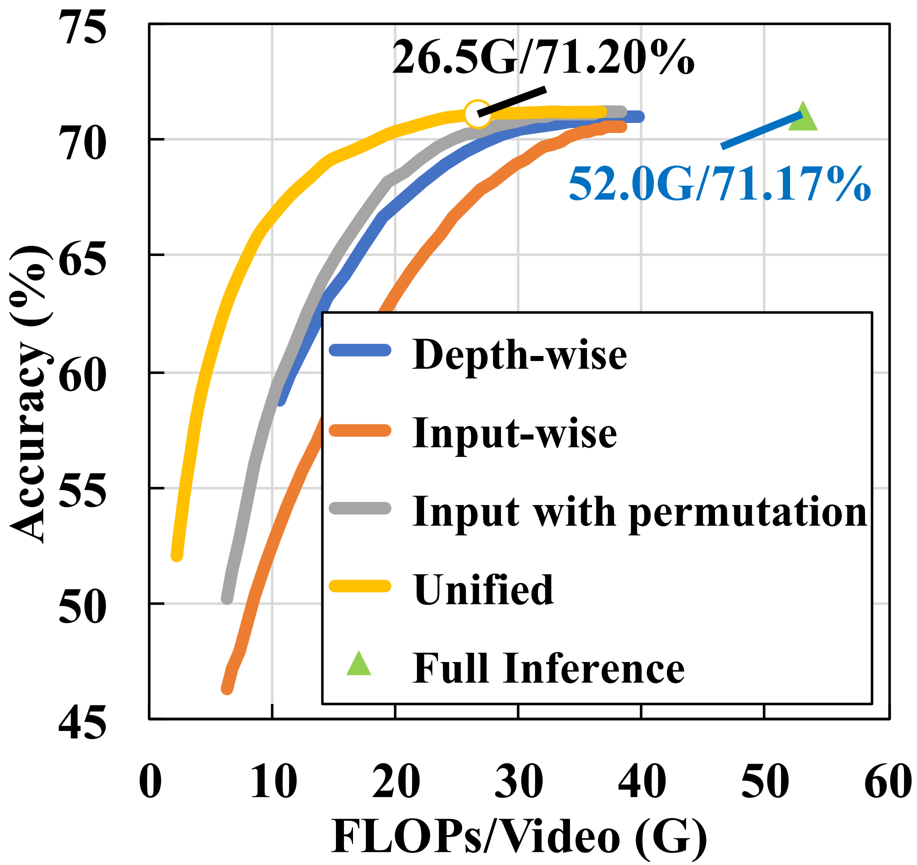
(b) With online temporal shift

Ablation experimental results with MSDNet backbone on Kinetics-400. "Full Inference" means that, for each video, only the prediction head of the last checkpoint is used.

Experiment: Ablation Study



(a) W/o online temporal shift



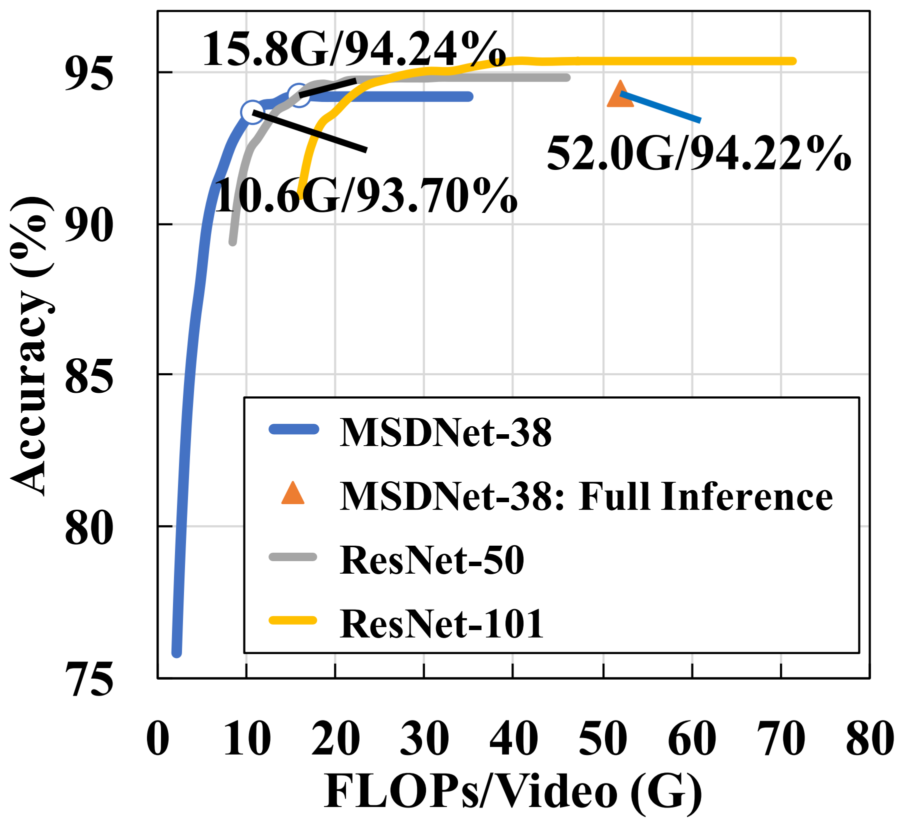
(b) With online temporal shift

FLOPs/Video ↓ ~ 53 %

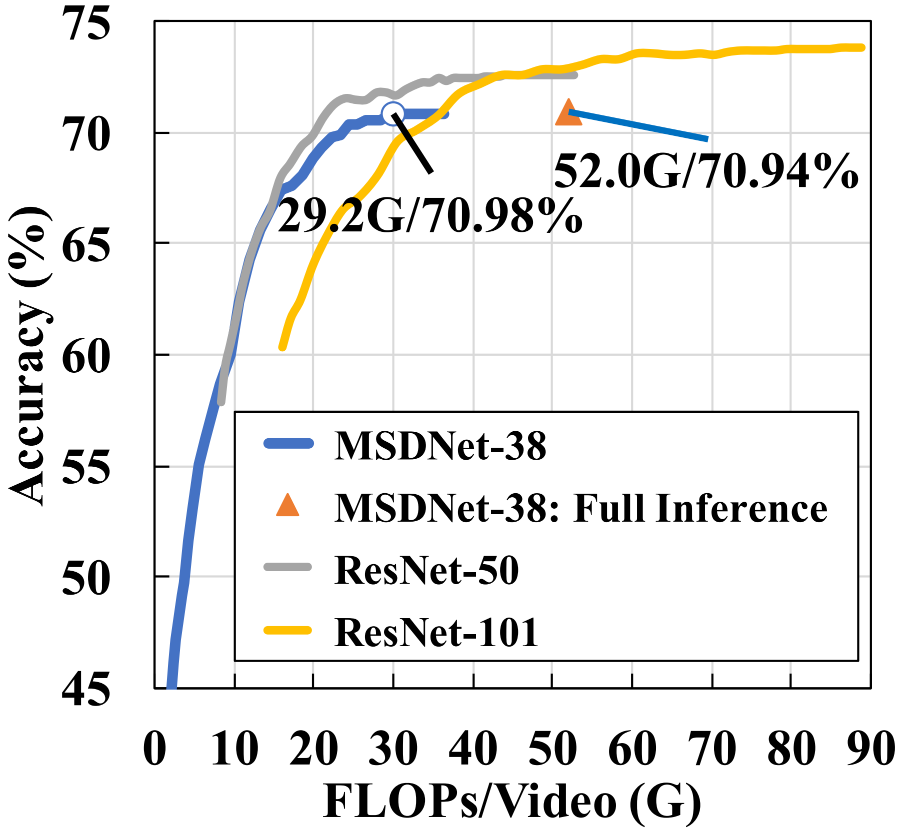
FLOPs/Video ↓ ~ 50 %

Experiment

CVPR SEATTLE WASHINGTON JUNE 16-18 2020



(a) UCF-101



(b) HMDB-51

FLOPs/Video ↓ ~ **70 %**

FLOPs/Video ↓ ~ **44 %**

Comparison with State-of-the-arts

CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

Kinetics-400

Framework	Backbone	Input × # Clips	Prec@1	# Params	FLOPs/Video
I3D [1]	3D BN-Inception	[All×3×256×256]×1	70.24	12.7M	544.44G
S3D [34]	3D BN-Inception	[All×3×224×224]×1	72.20	8.8M	518.6G
ARTNet with TSN [28]	3D ResNet-18	[16×3×112×112]×250	69.2	35.2M	5925G
MF-Net [2]	-	[16×3×224×224]×1 [16×3×224×224]×50	65.00 72.80	8.0M	11.1G 555G
ECO [35]	BN-Inception+3D ResNet-18	[16×3×224×224]×1	69.00	47.5M	64G
R(2+1)D RGB [27]	ResNet-34	[32×3×112×112]×10	72.00	63.8M	1524G
Nonlocal-I3d [31]	ResNet-50	[128×3×224×224]×1 [128×3×224×224]×30	67.30 76.50	35.33M	145.7G 4371G
TSN RGB [30]	BN-Inception ResNet-50 ResNet-50	[25×3×112×112]×10 [8×3×224×224]×1 [16×3×224×224]×1	69.1 66.80 67.80	10.7M 24.3M 24.3M	500G 33G 64G
TSM [20]	ResNet-50	[8×3×224×224]×1 [16×3×224×224]×1	70.60 72.50	24.3M 24.3M	33G 64G
StNet [12]	ResNet-50 ResNet-101	[25×15×256×256]×1 [25×15×256×256]×1	69.85 71.38	33.16M 52.15M	189.29G 310.50G
Proposed	MSDNet-38 (Full) MSDNet-38 ResNet-50 ResNet-101	[16×3×224×224]×1 [16×3×224×224]×1 [16×3×224×224]×1 [16×3×224×224]×1	71.17 **71.20** **72.57** **74.70**	62.31M 62.31M 29.12M 48.12M	52G **26.5G** **35G** **66G**

UCF-101 & HMDB-51

Method	Backbone	FLOPs	UCF-101	HMDB-51
ARTNet with TSN	3D ResNet-18	5925G	94.3	70.9
ECO	BNInception+ 3D ResNet-18	64G	92.8	68.5
I3D RGB	3D Inception-v1	544G	95.1	74.3
TSN RGB TSN_{8F} TSN_{16F}	BNInception ResNet-50	500G 33G 64G	91.1 91.5 91.4	- 63.2 63.6
TSM_{8F} TSM_{16F}	ResNet-50	33G 64G	94.0 94.5	70.3 70.7
StNet	ResNet-50	53G	93.5	-
Proposed	MSDNet-38	**15.8G** **29.2G**	**94.2** -	- **70.1**
	ResNet-50	**18.5G** **34.4G**	**94.7** -	- **72.34**
	ResNet-101	**34.6G** **69.1G**	**95.3** -	- **73.48**

Sth-Sth V1 & V2

Method	Backbone	Pretrain	FLOPs/Video	Something-Something v1		Something-Something v2			
				top-1 val	top-5 val	top-1 val	top-5 val	top-1 test	top-5 test
ECO_{16F} ECO_{EN} *Lite*	BNInception+ 3D ResNet-18	Kinetics	64G 267G	41.4 46.4	- -	- -			
I3D Non-local I3D+GCN	3D ResNet-50	Kinetics	306G 606G	41.6 46.1	72.2 76.8	- -	- -	- -	- -
TSN_{8F} TSN_{16F}	ResNet-50	Kinetics	33G 65G	19.7 19.9	46.6 47.3	27.8 30.0	57.6 60.5	- -	- -
TRN Multiscale TRN Two-Stream	BNInception	ImageNet	33G -	34.4 42.0	- -	48.8 55.5	77.6 83.1	50.9 56.2	79.3 83.2
TSM_{8F} TSM_{16F}	ResNet-50	Kinetics	33G 65G	43.4 44.8	73.2 74.5	58.2 58.7	84.8 84.8	- 59.9	- 85.9
Proposed	ResNet-50 MSDNet-38	ImageNet	**52.8(v1)/48.0G(v2)** **38.4G(v1)/35.4G(v2)**	**45.2** **46.5**	**75.2** **75.6**	**58.2** **60.0**	**85.2** **86.2**	- **60.1**	- **86.6**

Visualization

Playing badminton

Bench pressing

(a) Two video instances which stop at the first checkpoint

Visualization

Drawing

Garbage collecting

(b) Two video instances which stop at the middle checkpoint

Visualization

Trimming trees

Tobogganing

(c) Two video instances which stop at the last checkpoint

Thank you!

Dynamic Inference: A New Approach Toward Efficient Video Action Recognition

Contact: Wenhao Wu
wuwenhao17@mails.ucas.edu.cn
http://whwu95.github.io