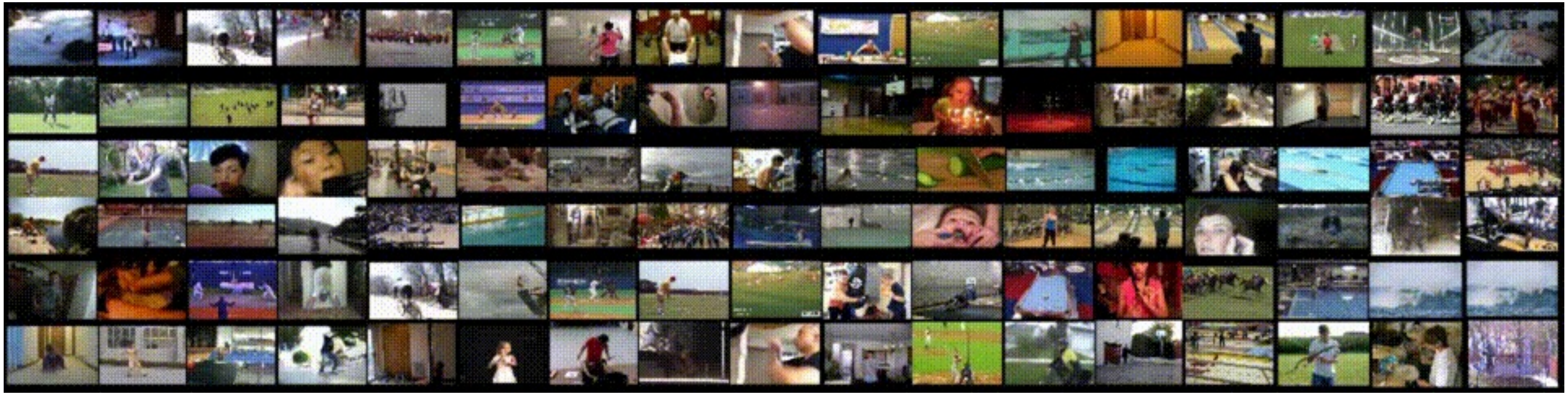
DSANet: Dynamic Segment Aggregation Network for Video-Level Representation Learning

Wenhao Wu[1*†], Yuxiang Zhao[1,2*], Yanwu Xu[3], Xiao Tan[1], Dongliang He[1],
Zhikang Zou[1], Jin Ye[1], Yingying Li[1], Mingde Yao[1], Zichao Dong[1], Yifeng Shi[1]
[1] Baidu Inc. [2] Shenzhen Institute of Advanced Technology, CAS [3] University of Pittsburgh

ACMMM 2021

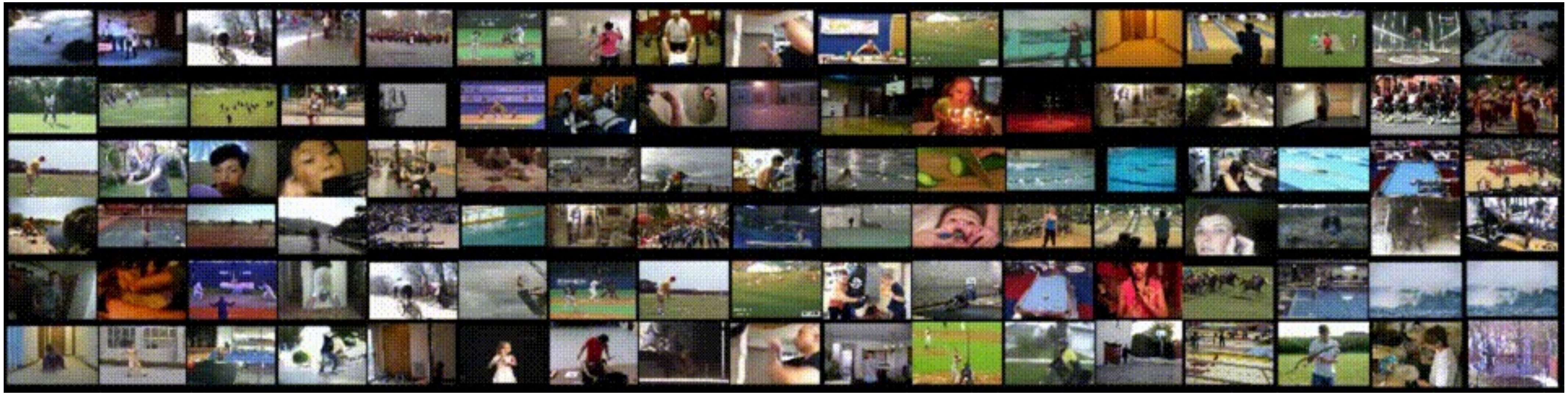
Task

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



Task

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



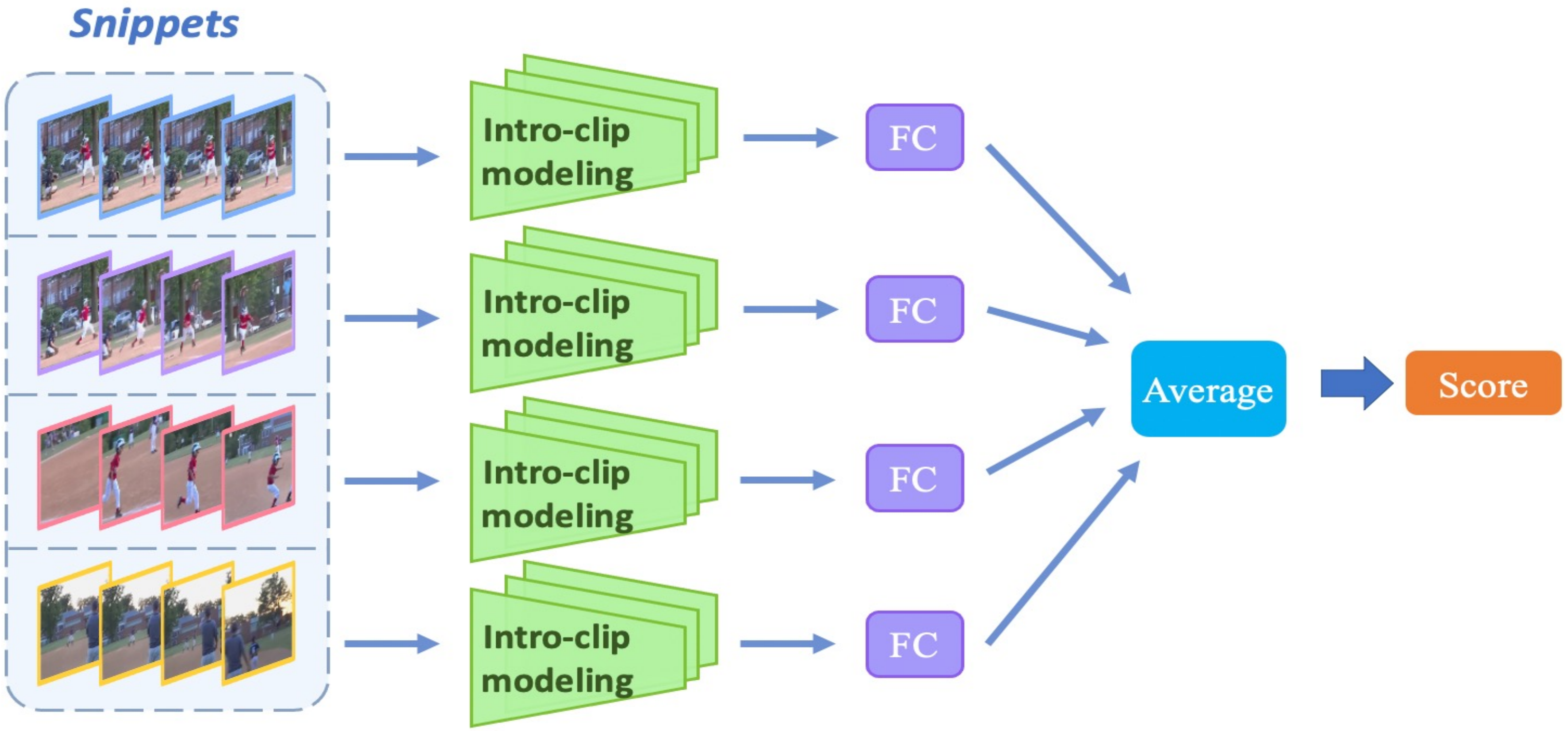
- More than simply recognizing objects
- Complex person-person interaction & people-object interactions
- Videos bring motions

Motivation

How to get the video-level prediction?

Classical Pipeline:

1. (**Training**) Intra-clip modeling (*e.g.*, C3D/TSM/SlowFast, *etc*)
2. (**Inference**) Average the predictions of multiple clips



Motivation

How to get the video-level prediction?

Classical Pipeline

Prominent problems:

- *No interaction among clips*
- *Training and Inference are not consistent*

*Can we learn **video-level** representations **directly**?*

Motivation

How to get the video-level prediction?

*Can we learn **video-level** representations **directly**?*

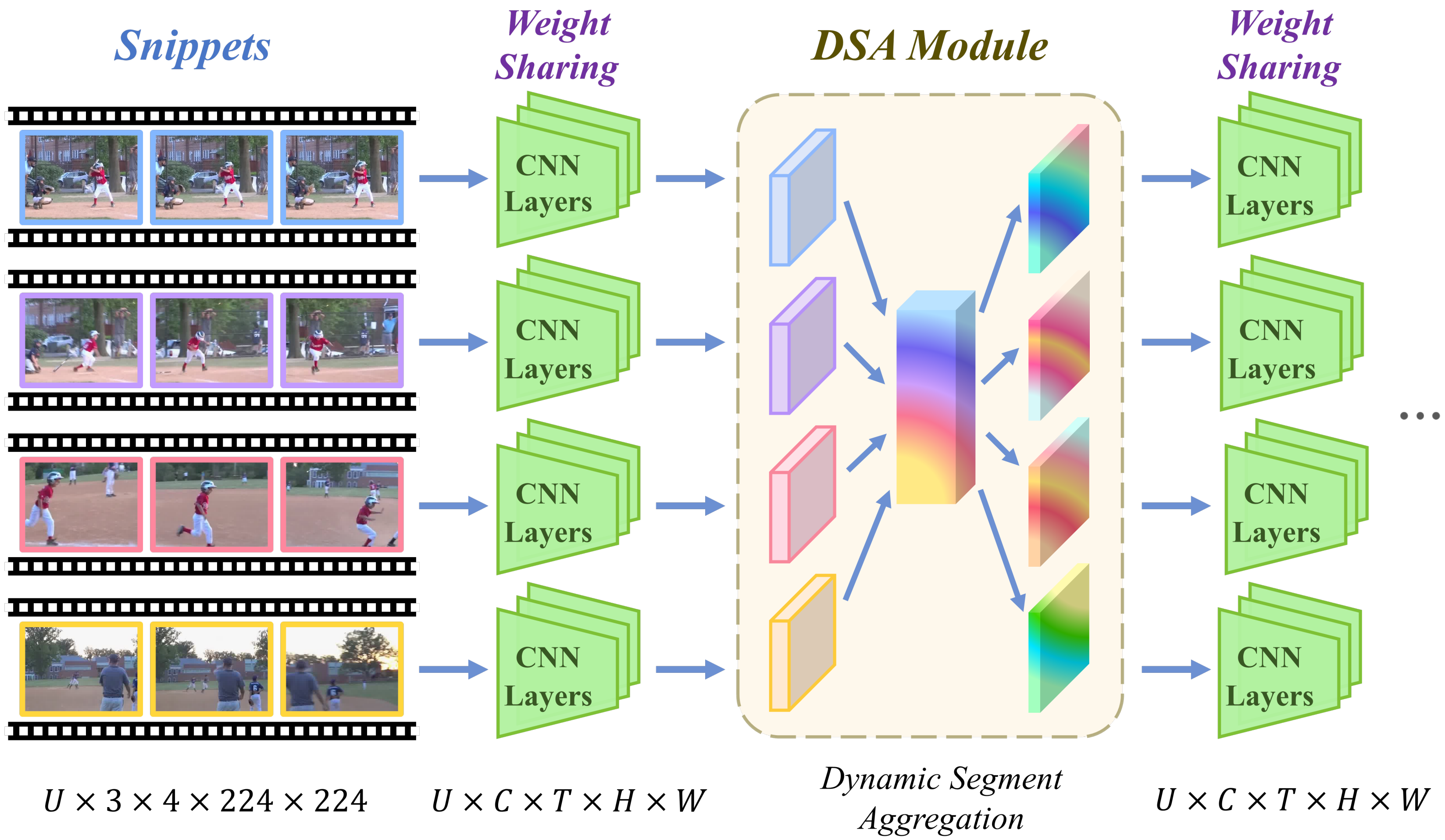
Intra-clip modeling (3D: T*H*W) → **Inter**-clip modeling (4D: U*T*H*W)

- TSN[1]: temporal modeling unexplored but simple
- 4D Convolution[2]: effective but expensive
- We focus on **efficient** and **effective** video-level representation learning

[1] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In Proc. ECCV.
[2] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. 2020. V4D: 4D Convolutional Neural Networks for Video-level Representation Learning. In Proc. ICLR.

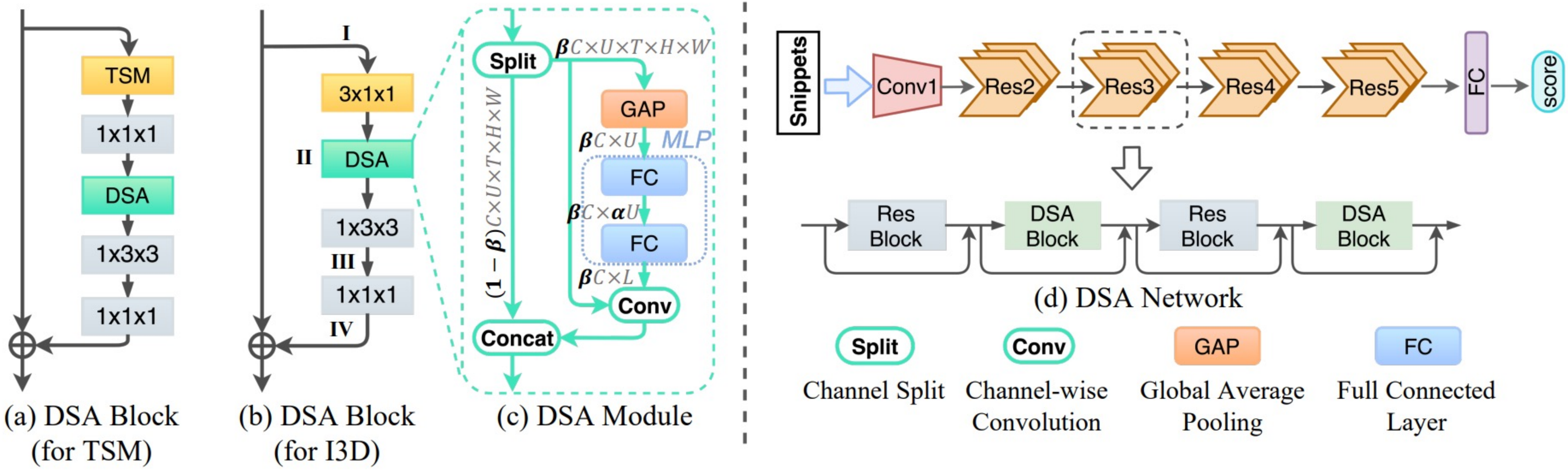
DSA Module

We propose a light-weight Dynamic Snippets Aggregation module to improve performance !



ACM multimedia
Chengdu, China OCT 20-24 2021

DSANet



Solving Problems

- Adaptively aggregate snippets to enhance temporal interaction
- Convolution on channel wise to reduce computation burden

Ablation Studies

(a) Study on the effectiveness of DSA module. T denotes the number of frames sampled from each video snippet, U denotes the number of snippets. Backbone: I3D R18.

Model	$T_{train} \times U_{train}$	$T_{infer} \times U_{infer} \times$#crop	Top-1	Top-5	Params
I3D R18	4×1	$4 \times 10 \times 3$	72.2	91.2	32.3M
I3D R18	16×1	$16 \times 10 \times 3$	73.4	91.1	32.3M
TSN+I3D R18	4×4	$4 \times 10 \times 3$	73.0	91.3	32.3M
V4D+I3D R18	4×4	$4 \times 10 \times 3$	75.6	92.7	33.1M
DSA+I3D R18	4×4	$4 \times 8 \times 3$	**77.3**	**93.9**	32.3M

Effectiveness

(i) Training FLOPs. Comparison with V4D, the extra computation cost brought by the DSA module is close to zero.

Model	Input size	FLOPs
TSN+I3D R50	$4 \times 4 \times 224^2 \times 3$	83.8G
V4D+I3D R50	$4 \times 4 \times 224^2 \times 3$	143.0G
DSA+I3D R50	$4 \times 4 \times 224^2 \times 3$	83.8G

Efficiency

Ablation Studies

(b) Study on different position to insert DSA module. Setting: I3D R50, α=2, β=1, stage: res_5.

Position	Top-1	Top-5
I	81.4	**95.4**
II	**81.5**	95.2
III	80.8	95.2
IV	81.4	95.1

(c) Parameter choices of α. Setting: I3D R50, Position II, β=1, inserted stage: res_5.

Setting	Top-1	Top-5
α=1	81.0	95.1
α=2	**81.5**	**95.2**
α=4	81.2	95.0
α=8	81.3	95.0

(d) The DSA blocks in different stage of I3D R50. Setting: Position II, α=2, β=1.

Stage	Top-1	Top-5
res{2}	81.4	94.7
res{3}	81.3	95.1
res{4}	81.3	**95.3**
res{5}	**81.5**	95.2

(e) Parameter choices of β. Setting: I3D R50, Position II, α=2, inserted stage: res_5.

Setting	Top-1	Top-5
β=1	81.5	95.2
β=1/2	**81.7**	**95.4**
β=1/4	81.6	95.0
β=1/8	81.5	95.0

(f) The number of DSA block inserted into I3D R50. Setting: Position II, α=2, β=1/8.

Stages	Blocks	Top-1	Top-5
res{5}	1	81.5	95.0
res{4,5}	4	81.5	95.3
res{3,4}	5	**81.8**	**95.4**
res{2,3}	3	81.4	95.1

Ablation Studies

Complementary with clip-based methods

(g) Different short-term temporal structure for DSA module.

Model	Top-1	Top-5
TSM R50	77.4	93.4
DSA+TSM R50	**80.4**	**95.0**
I3D R50	78.0	93.9
DSA+I3D R50	**81.8**	**95.4**

Complementary with different backbones

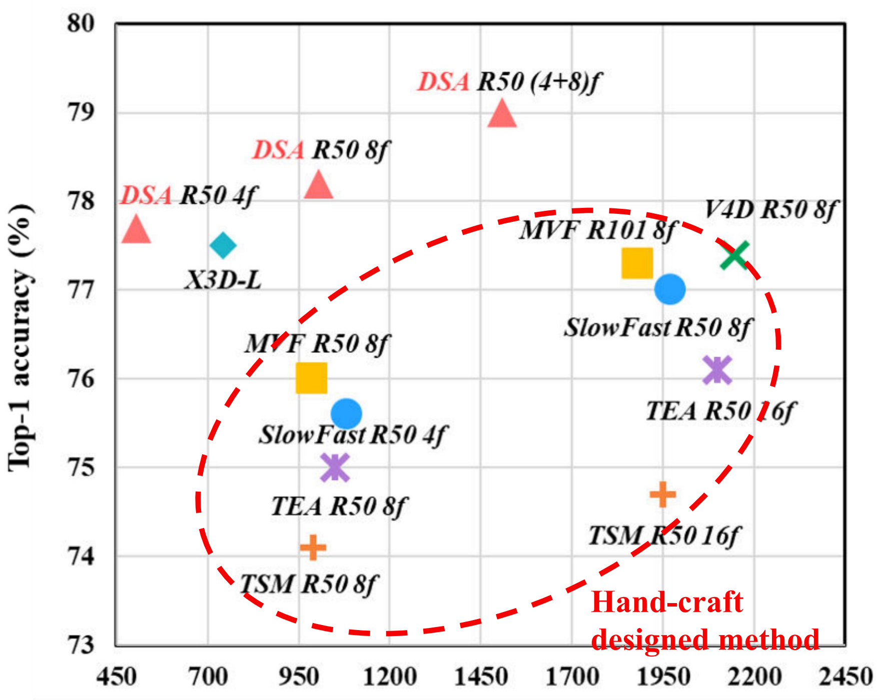
(h) Study on the effectiveness of DSA module with different backbones (I3D R18, I3D R50). SENet+I3D uses SE module to replace the DSA module in DSANet.

Arch.	I3D	SENet+I3D	DSA+I3D
ResNet18	72.2	73.8	**77.3**
ResNet50	78.0	78.5	**81.8**

Comparison with SOTAs

ACM multimedia
Chengdu, China OCT 20-24 2021

Kinetics-400



Accuracy-computation trade-off

Method	Backbone	$T_{infer} \times U_{infer} \times$ #crop	GFLOPs	Top-1	Top-5
TSM [17]	ResNet-50	8×10×3	33×30=990	74.1%	91.2%
TEINet [19]	ResNet-50	8×10×3	33×30=990	74.9%	91.8%
TEA [16]	ResNet-50	8×10×3	35×30=1050	75.0%	91.8%
TANet [20]	ResNet-50	8×10×3	43×30=1290	76.1%	92.3%
MVFNet [34]	ResNet-50	8×10×3	33×30=990	76.0%	92.4%
NL+I3D [31]	3D ResNet-50	32×10×3	70.5×30=2115	74.9%	91.6%
NL+I3D [31]	3D ResNet-50	128×10×3	282×30=8460	76.5%	92.6%
X3D-L [7]	-	16×10×3	24.8×30=744	77.5%	92.9%
Slowfast [8]	3D R50+3D R50	(4+32)×10×3	36.1×30=1083	75.6%	92.1%
Slowfast [8]	3D R50+3D R50	(8+32)×3×10	65.7×30=1971	77.0%	92.6%
Slowfast [8]	3D R101+3D R101	(8+32)×3×10	106×30=3180	77.9%	93.2%
Slowonly [8]	3D ResNet-50	8×10×3	41.9×30=1257	74.9%	91.5%
V4D+I3D [39]	3D ResNet-50	8×10×3	286.1×2.5×3=2146*	77.4%	93.1%
DSA+I3D (Ours)	3D ResNet-50	4×8×3	83.8×2×3=503	**77.7%**	**93.1%**
DSA+I3D (Ours)	3D ResNet-50	8×8×3	167.7×2×3=1006	**78.2%**	**93.2%**
DSA+I3D (Ours)	3D ResNet-50	(4+8)×8×3	251.5×2×3=1509	**79.0%**	**93.7%**

Comparison with SOTAs

ACM multimedia
Chengdu, China OCT 20-24 2021

Mini-Kinetics-200

Method	Backbone	$T_{train} \times U_{train}$	$T_{infer} \times U_{infer} \times$ #crop	Top-1	Top-5
S3D [37]	S3D Inception	64×1	N/A	78.9%	-
I3D [38]	3D ResNet50	32×1	$32 \times 10 \times 3$	75.5%	92.2%
I3D [38]	3D ResNet101	32×1	$32 \times 10 \times 3$	77.4%	93.2%
I3D+NL [38]	3D ResNet50	32×1	$32 \times 10 \times 3$	77.5%	94.0%
I3D+CGNL [38]	3D ResNet50	32×1	$32 \times 10 \times 3$	78.8%	94.4%
I3D+NL [38]	3D ResNet101	32×1	$32 \times 10 \times 3$	79.2%	93.2%
I3D+CGNL [38]	3D ResNet101	32×1	$32 \times 10 \times 3$	79.9%	93.4%
V4D+I3D [39]	3D ResNet18	4×4	$4 \times 10 \times 3$	75.6%	92.7%
V4D+I3D [39]	3D ResNet50	4×4	$4 \times 10 \times 3$	80.7%	95.3%
DSA+I3D (Ours)	3D ResNet18	4×4	$4 \times 8 \times 3$	77.3%	93.9%
DSA+I3D (Ours)	3D ResNet50	4×4	$4 \times 8 \times 3$	**81.8%**	**95.4%**

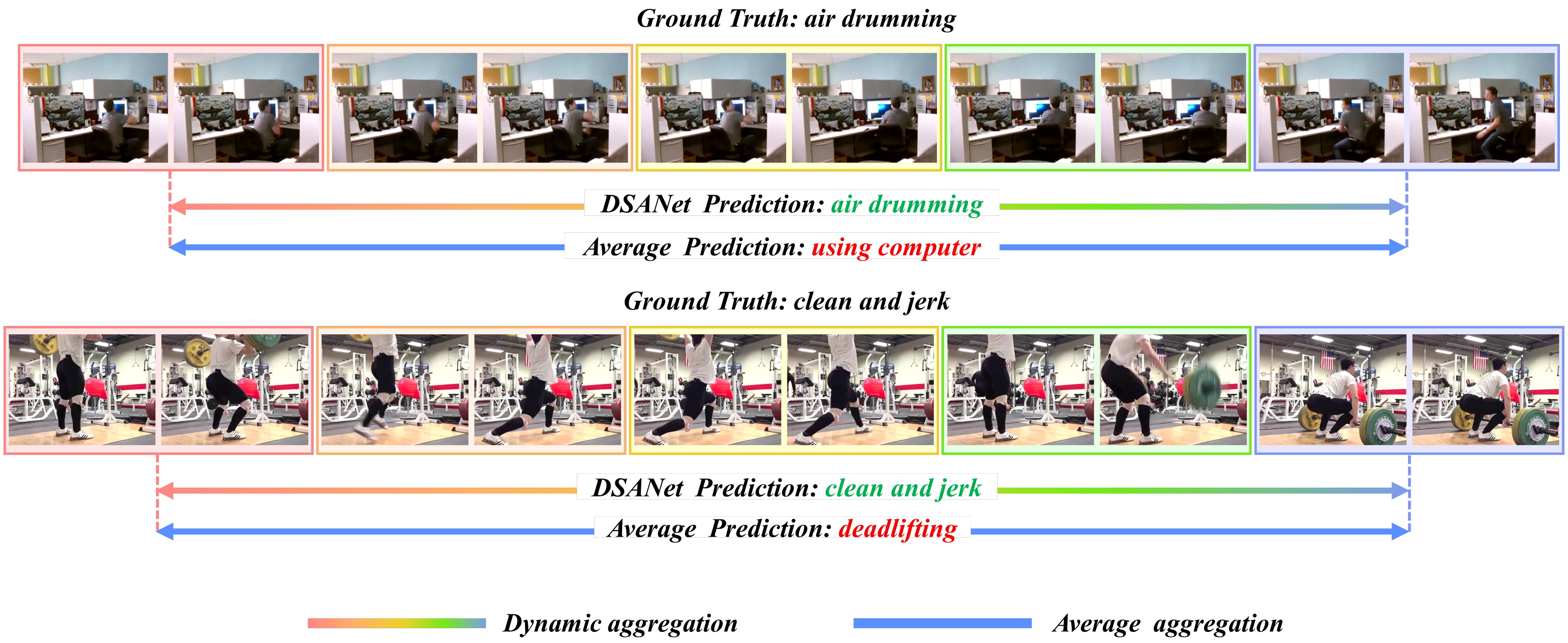
ActivityNet v1.3

Model	Backbone	mAP
TSN [30]	BN-Inception	79.7%
TSN [30]	Inception V3	83.3%
TSN-Top3 [30]	Inception V3	84.5%
V4D+I3D [39]	3D ResNet50	88.9%
DSA+I3D (Ours)	3D ResNet50	**90.5%**

Something-Something V1

Method	Backbone	Top-1
MultiScale TRN [40]	BN-Inception	34.4%
ECO [41]	BN-Inception+3D ResNet 18	46.4%
S3D-G [37]	S3D Inception	45.8%
Nonlocal+GCN [32]	3D ResNet50	46.1%
TSM [17]	ResNet50	47.2%
I3D (our impl.)	3D ResNet50	48.7%
V4D+I3D [39]	3D ResNet50	50.4%
DSA+I3D (Ours)	3D ResNet50	**51.8%**

Visualization



Baidu 百度

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

ACM multimedia
Chengdu, China OCT 20-24 2021

Thank you for your attention!

- Codes

https://github.com/whwu95/DSANet

- Contact

Wenhao Wu
Baidu Inc.
wuwenhao17@mails.ucas.edu.cn