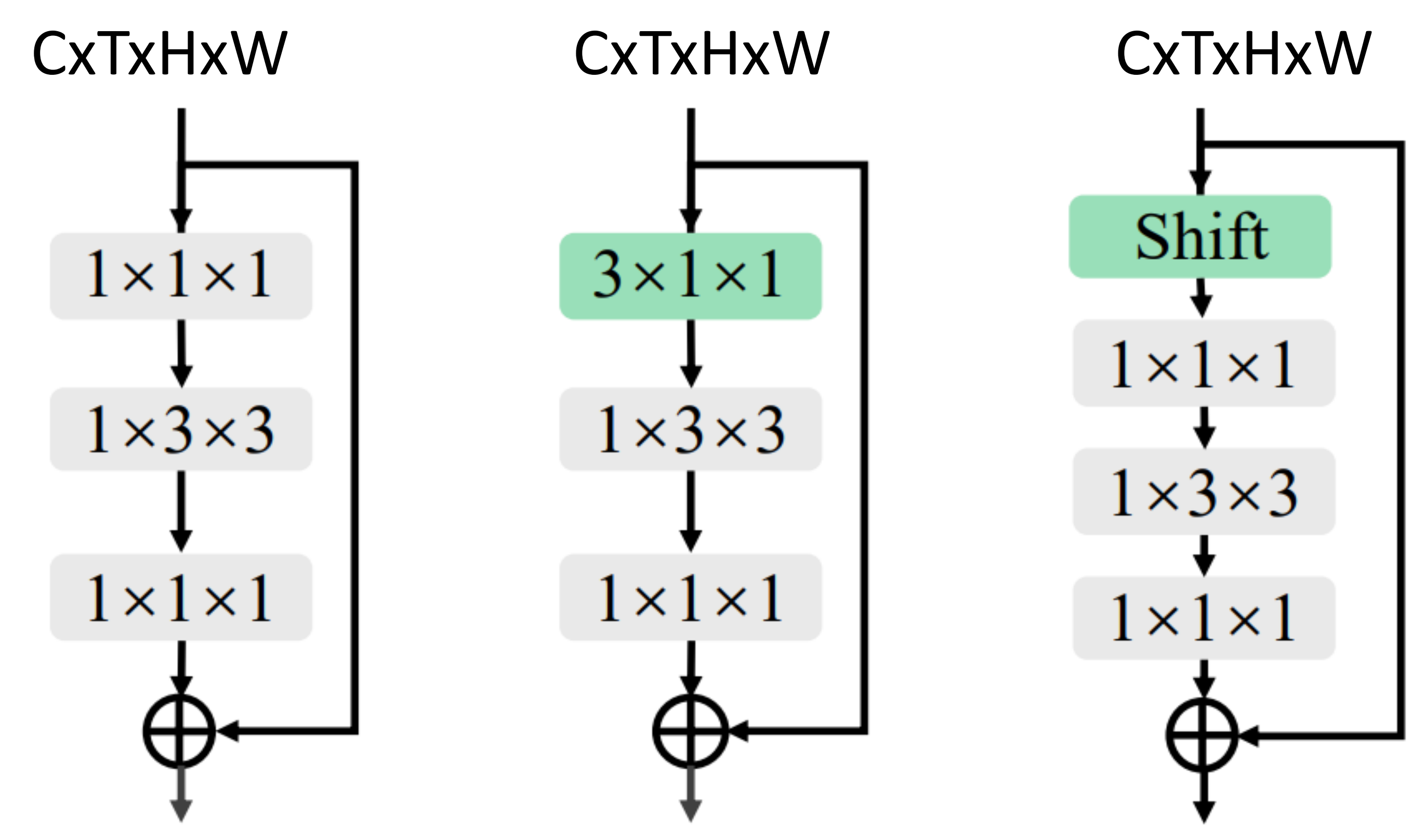
MVFNet: Multi-View Fusion Network for Efficient Video Recognition

Wenhao Wu[1], Dongliang He[1], Tianwei Lin[1], Fu Li[1], Chuang Gan[2], Errui Ding[1]

[1]Department of Computer Vision Technology (VIS), Baidu Inc.
[2]MIT-IBM Watson AI Lab

MOTIVATION



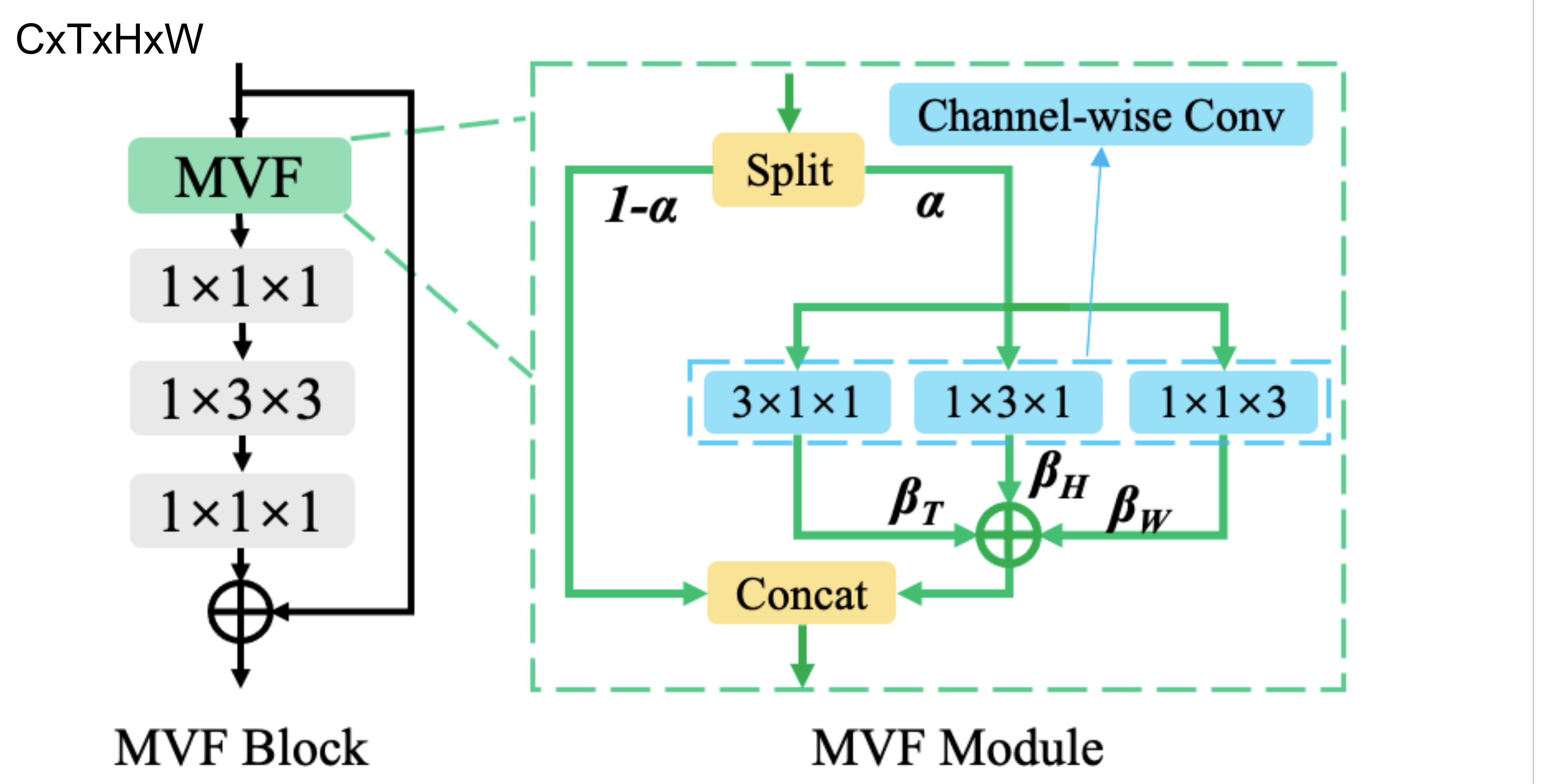
(a) C2D (b) SlowOnly (c) TSM

- Efficient spatiotemporal modeling is the key to video recognition
- Classical C2D: temporal modeling unexplored but simple
- 3D CNN, *e.g.*, SlowFast or SlowOnly: effective but expensive
- TSM enables C2D to model temporal relationship at nearly zero cost
  - ***fixed*** channel-wise 3x1x1 conv: kernel of [0,0,1] for forward shift and [1,0,0] for backward shift

CONTRIBUTION

- ➢ Instead of only temporal modeling, we propose to exploit dynamic inside the three dimensional video signal from multiple viewpoints. A novel MVF module is designed to better exploit spatiotemporal dynamics.
- ➢ The MVF module works in a plug-and-play way and can be integrated easily with existing 2D CNN backbones. Our MVFNet is a generalized video modeling network and it can specialize to become recent state-of-the-arts.
- ➢ Extensive experiments on five public benchmark datasets demonstrate that the proposed MVFNet outperforms the state-of-the-art methods with computational cost (GFLOPs) comparable to 2D CNN.

METHOD



MVFNet = $\underbrace{\text{Learnable Temporal “Shift”}}_{\textit{Viewpoint of HW-T}}$ + $\underbrace{\text{Learnable Horizontal “Shift”}}_{\textit{Viewpoint of WT-H}}$ + $\underbrace{\text{Learnable Vertical “Shift”}}_{\textit{Viewpoint of TH-W}}$

EXPERIMENTS

Method	Backbone	Frames×Crops×Clips	GFLOPs	Top-1	Top-5
I3D (Carreira et al. 2017)	Inception V1	64×N/A×N/A	108×N/A	72.1%	90.3%
S3D-G (Xie et al. 2018)	Inception V1	64×3×10	71.4×30	74.7%	93.4%
TSN (Wang et al. 2016)	Inception V3	25×10×1	80×10	72.5%	90.2%
ECO-RGB$_{En}$ (Zolfaghari et al. 2018)	BNIncep+Res3D-18	92×1×1	267×1	70.0%	-%
R(2+1)D (Tran et al. 2018)	ResNet-34	32×1×10	152×10	74.3%	91.4%
X3D-M (Feichtenhofer 2020)	-	16×3×10	6.2×30	76.0%	92.3%
STM (Jiang et al. 2019)	ResNet-50	16×3×10	67×30	73.7%	91.6%
TSM (Lin, Gan, and Han 2019)	ResNet-50	8×3×10	33×30	74.1%	91.2%
SlowOnly (Feichtenhofer et al. 2019)	ResNet-50	8×3×10	41.9×30	74.9%	91.5%
TEINet (Liu et al. 2020)	ResNet-50	8×3×10	33×30	74.9%	91.8%
TEA (Li et al. 2020b)	ResNet-50	8×3×10	33×30	75.0%	91.8%
Slowfast (Feichtenhofer et al. 2019)	R50+R50	(4+32)×3×10	36.1×30	75.6%	92.1%
NL+I3D (Wang et al. 2018b)	ResNet-50	32×3×10	70.5×30	74.9%	91.6%
NL+I3D (Wang et al. 2018b)	ResNet-50	128×3×10	282×30	76.5%	92.6%
MVFNet	ResNet-50	8×3×10	32.9×30	**76.0%**	**92.4%**
MVFNet	ResNet-50	16×3×10	65.8×30	**77.0%**	**92.8%**
ip-CSN (Tran et al. 2019)	ResNet-101	32×3×10	82×30	76.7%	92.3%
SmallBig (Li et al. 2020a)	ResNet-101	32×3×4	418×12	77.4%	93.3%
SlowOnly (Feichtenhofer et al. 2019)	ResNet-101	16×3×10	185×30	77.2%	-%
NL+I3D (Wang et al. 2018b)	ResNet-101	128×3×10	359×30	77.7%	93.3%
Slowfast (Feichtenhofer et al. 2019)	R101+R101	(8+32)×3×10	106×30	77.9%	93.2%
Slowfast (Feichtenhofer et al. 2019)	R101+R101	(16+64)×3×10	213×30	**78.9%**	**93.5%**
TPN (Yang et al. 2020)	ResNet-101	32×3×10	374×30	**78.9%**	**93.9%**
MVFNet	ResNet-101	8×3×10	62.7×30	77.4%	92.9%
MVFNet	ResNet-101	16×3×10	125.4×30	78.4%	93.4%
MVFNet$_{En}$	R101+R101	(16+8)×3×10	188.1×30	**79.1%**	**93.8%**

Table. Comparison with the state-of-the-art models on Kinetics-400 dataset.

EXPERIMENTS

Setting	Sth-sth v1				Kinetics-400			
	#F	Top-1	Top-5	FLOPs	#F	Top-1	Top-5	FLOPs
α=0	8	17.12	43.46	32.88G	4	71.87	90.02	16.44G
α=1/8	8	49.74	78.09	32.90G	4	**74.21**	91.34	16.45G
α=1/4	8	49.24	77.91	32.92G	4	74.18	**91.46**	16.46G
α=1/2	8	**50.48**	**79.14**	32.96G	4	74.21	91.42	16.48G
α=1	8	49.73	77.94	33.04G	4	73.75	91.40	16.52G

(a) **Parameter choices of α**. Backbone: R-50.

Stages	Blocks	Sth-sth v1, α=1/2				Kinetics-400, α=1/8			
		#F	Top-1	Top-5	FLOPs	#F	Top-1	Top-5	FLOPs
None	0	8	17.12	43.46	32.88G	4	71.87	90.02	16.44G
res{5}	3	8	46.02	75.60	32.90G	4	73.46	91.09	16.44G
res{4,5}	9	8	**50.48**	**79.14**	32.96G	4	74.21	91.34	16.45G
res{3,4,5}	13	8	49.72	78.82	33.04G	4	74.08	91.51	16.46G
res{2,3,4,5}	16	8	49.95	77.96	33.12G	4	**74.22**	**91.56**	16.47G

(b) **The number of MVF Blocks** inserted into R-50.

Views	Sth v1		K400	
	#F	Top-1	#F	Top-1
T	8	49.13	4	73.72
T-H	8	49.22	4	74.01
T-W	8	49.31	4	73.88
T-H-W	8	**50.48**	4	**74.21**
T-H-W (S)	8	47.21	4	73.81

(c) **Study on the different views of MVF module**. Backbone: R-50. S denotes weight sharing.

Method	Sth v1	K400	FLOPs	Params
	Top-1	Top-1		
C2D	17.1	71.4	32.9G	24.3M
TSM	47.2	74.1	32.9G	24.3M
SlowOnly	-	74.9	41.9G	32.4M
CoST*	-	-	45.8G	24.3M
MVFNet	**50.5**	**76.0**	32.9G	24.3M

(d) **Study on the effectiveness of MVFNet**. Backbone: R-50, 8f input. * indicates our implementation.

	#F	Top-1	FLOPs
R-50	4	74.21	16.45G
	8	75.99	32.90G
	16	77.04	65.81G
R-101	4	75.98	31.36G
	8	77.46	62.72G
	16	78.42	125.45G

(e) **Advanced backbones for MVFNet on Kinetics-400**.

	Model	Top-1	FLOPs
Mb-V2	C2D	64.4	1.25G
	MVF	67.5	1.25G
R-50	C2D	71.9	16.44G
	MVF	74.2	16.48G

(f) **Different backbones for MVFNet on Kinetics-400**. Mb-V2 denotes MobileNet-V2.

Table. Ablation studies on Something-Something V1 and Kinetics-400.

Method	Backbone	Frames×Crops×Clips	FLOPs	Pre-train	V1 Val Top-1 (%)	V2 Val Top-1 (%)
I3D (Wang et al. 2018)	3D ResNet50	32×3×2	153G×3×2	ImageNet + K400	41.6	-
NL I3D (Wang et al. 2018)	3D ResNet50		168G×3×2		44.4	-
NL I3D+GCN (Wang et al. 2018)	3D ResNet50+GCN		303G×3×2		46.1	-
ECO (Zolfaghari et al. 2018)	BNIncep+3D Res18	8×1×1	32G×1×1	K400	39.6	-
ECO$_{En}$ (Zolfaghari et al. 2018)		92×1×1	267G×1×1		46.4	-
S3D-G (Xie et al. 2018)	Inception	64×1×1	71G×1×1	K400	48.2	-
TSN (Wang et al. 2016)	ResNet50	8×3×2	33G×3×2	ImageNet	20.5	30.4
TSM (Lin et al. 2019)	ResNet50	8×3×2	33G×3×2	ImageNet	47.2	61.2
		16×3×2	65G×3×2		48.4	63.1
STM (Jiang et al. 2019)	ResNet50	8×3×10	33G×3×10	ImageNet	49.2	62.3
		16×3×10	67G×3×10		50.7	64.3
TEINet (Liu et al. 2020)	ResNet50	8×3×10	33G×3×10	ImageNet	48.8	64.0
		16×3×10	66G×3×10		51.0	64.7
TEA (Li et al. 2020b)	ResNet50	8×3×10	35G×3×10	ImageNet	51.7	-
		16×3×10	70G×3×10		52.3	-
MVFNet	ResNet50	8×1×1	33G×1×1	ImageNet	48.8	60.8
		8×3×2	33G×3×2		50.5	63.5
		16×1×1	66G×1×1		51.0	62.9
		16×3×2	66G×3×2		**52.6**	**65.2**
		(16+8)×3×2	99G×3×2		**54.0**	**66.3**

Table. Comparison with Other Solutions on Sth-Sth-v1/v2.

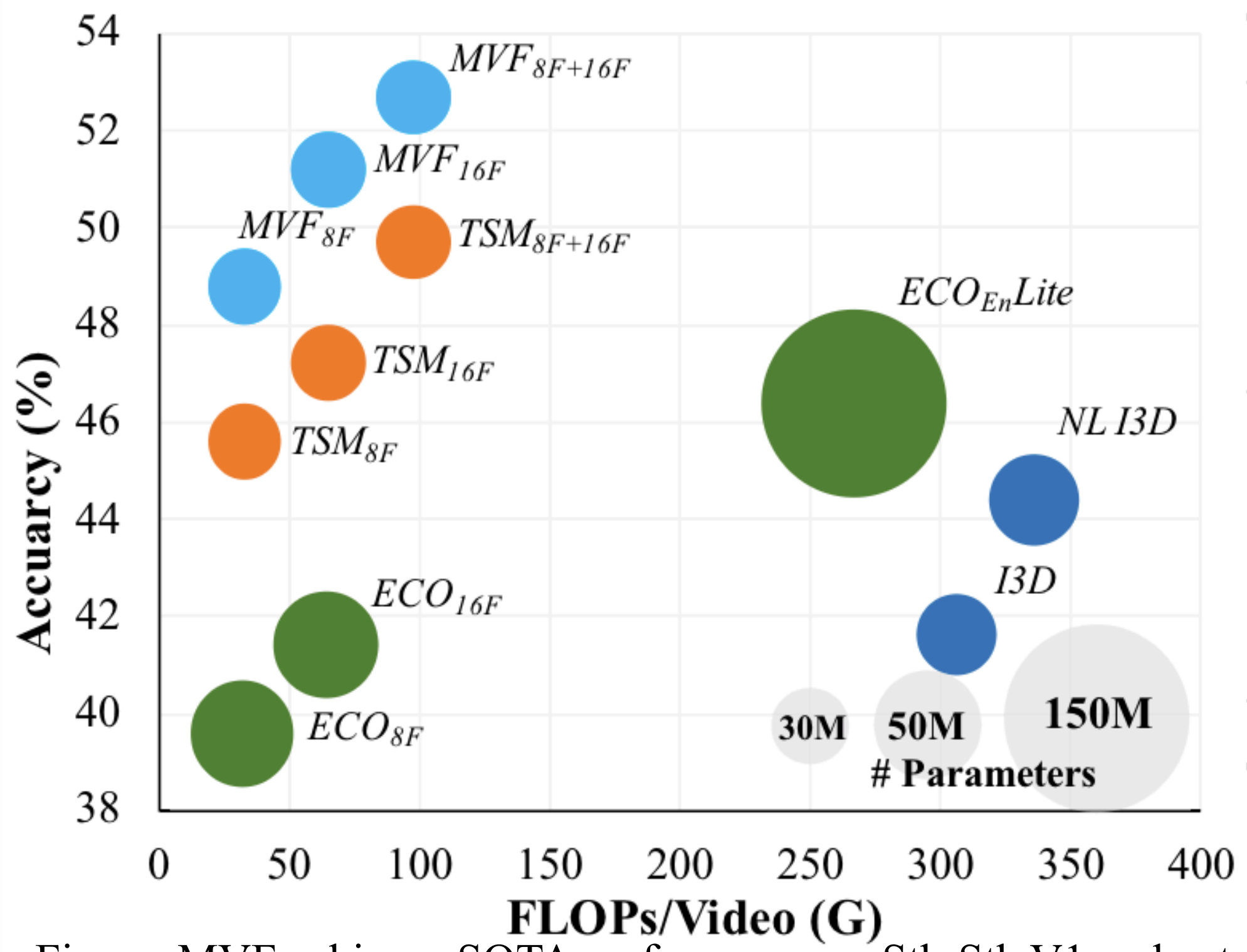


Figure. MVF achieves SOTA performance on Sth-Sth V1 and get better accuracy-computation trade-off than I3D, ECO and TSM.

Method	Backbone	UCF-101	HMDB-51
ECO$_{En}$	BNIncep+Res3D-18	94.8%	72.4%
ARTNet	ResNet-18	94.3%	70.9%
I3D	Inception V1	95.6%	74.8%
R(2+1)D	Inception V1	96.8%	74.5%
S3D-G	Inception V1	96.8%	75.9%
TSN	BNInception	91.1%	-
StNet	ResNet-50	93.5%	-
TSM	ResNet-50	95.9%	73.5%
STM	ResNet-50	96.2%	72.2%
TEINet	ResNet-50	96.7%	72.1%
MVFNet	ResNet-50	**96.6%**	**75.7%**

Table. **Mean class accuracy** on UCF-101 and HMDB-51 achieved by different methods which are transferred from their **Kinetics** models with RGB modality (over 3 splits).