Group Contextualization for Video Recognition

Yanbin Hao[1] Hao Zhang[2*] Chong-Wah Ngo[2] Xiangnan He[1]

[1]University of Science and Technology of China

[2]Singapore Management University

中国科学技术大学 University of Science and Technology of China

SMU SINGAPORE MANAGEMENT UNIVERSITY

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Code & Paper

https://github.com/haoyanbin918/Group-Contextualization

Goal and Contribution

Goal: Exploring various axial contexts to separately calibrate video feature channel groups in parallel with little computational overhead.

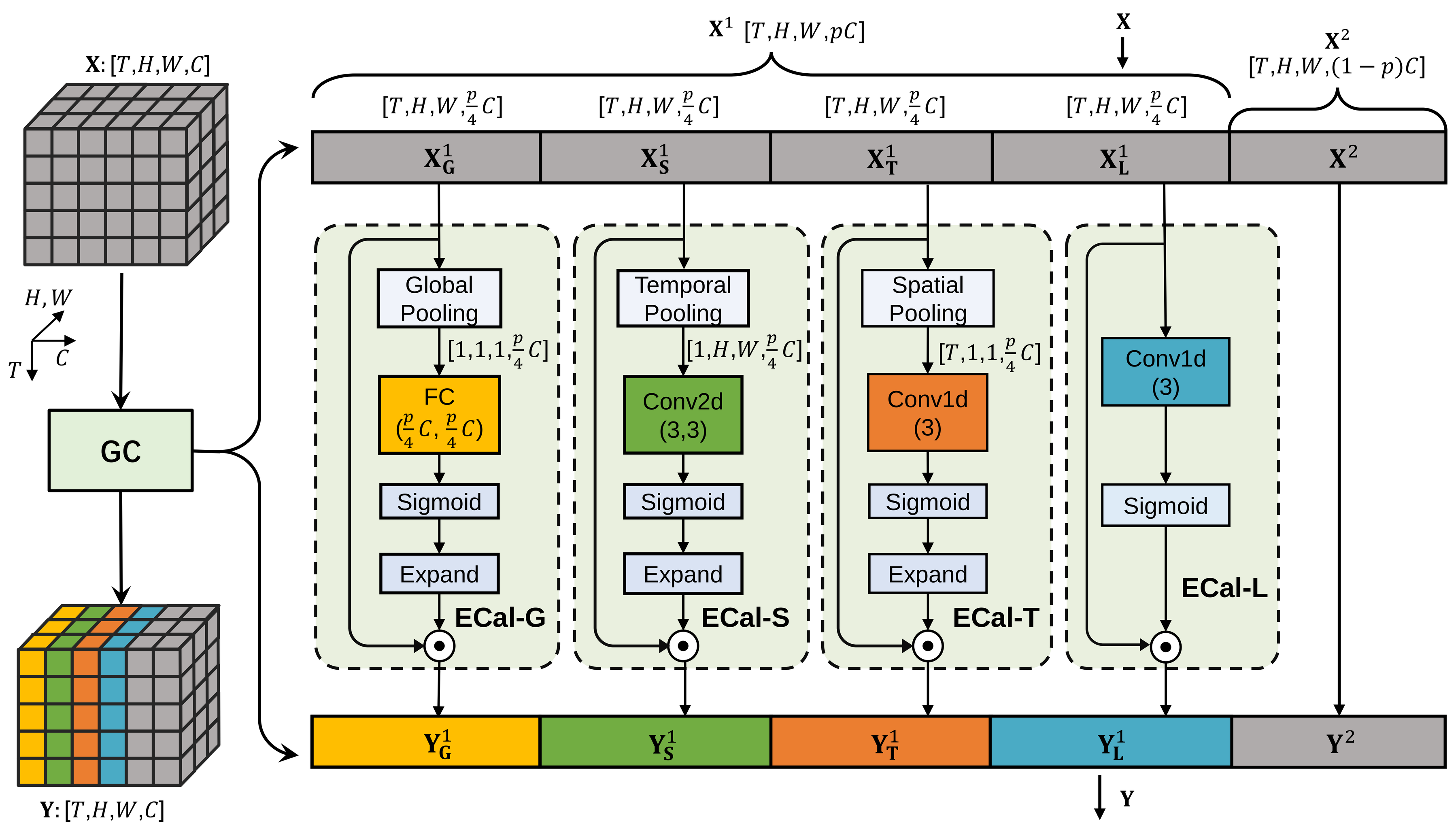
Contributions:

- We propose a new regime named group contextualization (GC) for video feature refinement, where a family of efficient element-wise calibrators (ECals) are purposely designed to model local and global axial contexts.
- The proposed GC module is easily integrated into various basic video CNNs without incurring significant computational burden and leads to notable performance gains.

Method

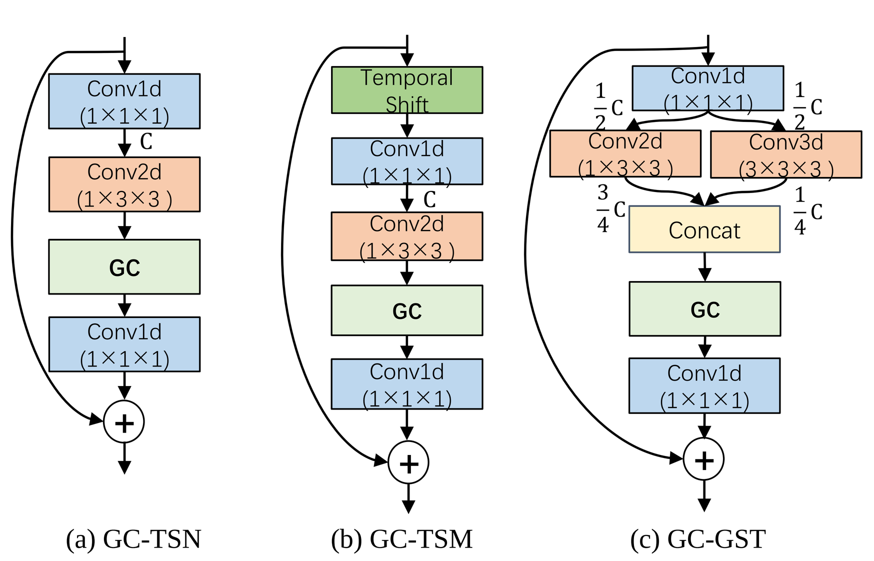
Module Architecture:

- The input CNN feature $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$ is firstly split into two channel groups $\mathbf{X}^1 \in \mathbb{R}^{T\times H\times W\times pC}$ and $\mathbf{X}^2 \in \mathbb{R}^{T\times H\times W\times (1-p)C}$. Then, four feature calibrators (ECal-G/S/T/L) are customized to focus on four different axial perspectives and separately refine the four feature channel subgroups ($\frac{p}{4}C$ channels) of $\mathbf{X}^1$ in parallel. All ECals share the similar cascaded structure of "GAP/None+FC/Conv+Sigmoid" for efficiency.



Integrated Networks:

- We integrate the GC module into three basic video networks, i.e., TSN, TSM, and GST, and a more advanced network, i.e., TDN, referred to as GC-TSN, GC-TSM, GC-GST and GC-TDN, respectively.
- We also empirically investigate a new position setting, i.e., the loop version, to examine the effect of channel position.



Channel Portion ID		0	1	2	3	4	5	6	7
Res1	Bottleneck-1	G	S	T	L				
	Bottleneck-2		G	S	T	L			
	Bottleneck-3			G	S	T	L		
Res2	Bottleneck-1				G	S	T	L	
	Bottleneck-2					G	S	T	L
	Bottleneck-3	L					G	S	T
	Bottleneck-4	T	L					G	S
Res3	Bottleneck-1	S	T	L					G
	Bottleneck-2	G	S	T	L				
	Bottleneck-3		G	S	T	L			
	Bottleneck-4			G	S	T	L		
	Bottleneck-5				G	S	T	L	
	Bottleneck-6					G	S	T	L
Res4	Bottleneck-1	L					G	S	T
	Bottleneck-2	T	L					G	S
	Bottleneck-3	S	T	L					G

Experiments & Results

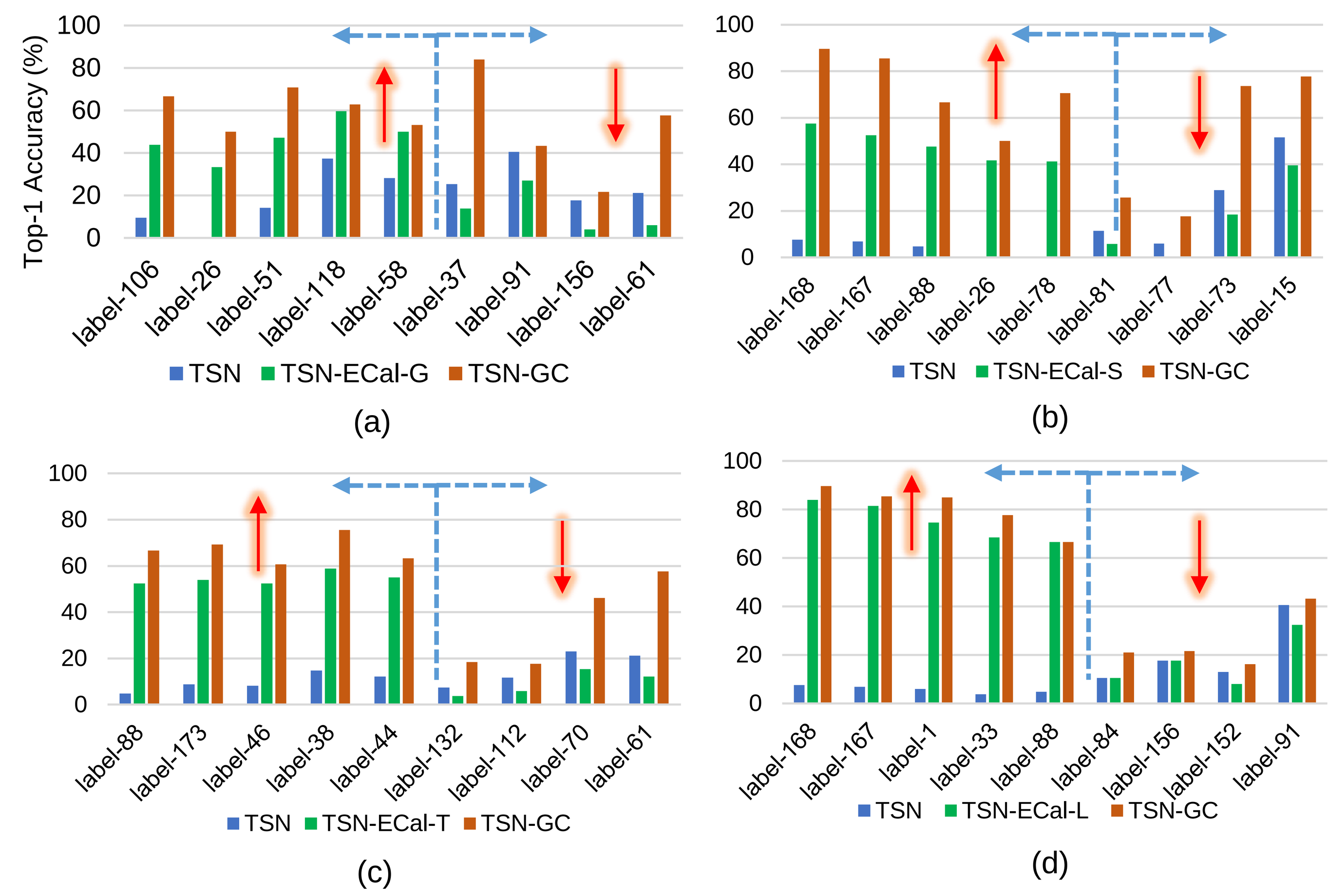
Dataset: We conduct experiments on several different benchmarks, for example, Something-Something V1&V2 and Kinetics-400 for video recognition.

Effectiveness of GC Module on Something-Something V1:

Backbone	Calibrator	(p, Channel)	Params	FLOPs	Top-1 (%)
TSN	—	—	23.9M	32.9G	19.7
	SE3D	—	26.4M	32.9G	27.8 (+8.1)
	GE3D-G	—	23.9M	32.9G	22.3 (+2.6)
	GE3D-C	—	25.2M	33.3G	44.2 (+24.5)
	S3D-G	—	25.1M	32.9G	28.0 (+8.3)
	NLN	—	31.2M	49.4G	30.3 (+10.6)
	ECal-G	$(\frac{1}{2}, \frac{1}{8}C)$	23.9M	32.9G	26.3 (+6.6)
		$(1, \frac{1}{4}C)$	23.9M	32.9G	27.3 (+7.6)
	ECal-T	$(\frac{1}{2}, \frac{1}{8}C)$	23.9M	32.9G	35.9 (+16.2)
		$(1, \frac{1}{4}C)$	24.1M	32.9G	36.4 (+16.7)
	ECal-S	$(\frac{1}{2}, \frac{1}{8}C)$	24.0M	32.9G	34.0 (+14.3)
		$(1, \frac{1}{4}C)$	24.6M	33.0G	34.1 (+14.4)
	ECal-L	$(\frac{1}{2}, \frac{1}{8}C)$	23.9M	33.0G	44.8 (+25.1)
		$(1, \frac{1}{4}C)$	24.1M	33.2G	44.9 (+25.2)
	GC	$(\frac{1}{2}, \frac{1}{2}C)$	24.2M	33.0G	47.1 (+27.4)
		$(1, C)$	25.1M	33.3G	47.9 (+28.2)
		$(1, C)$, loop	25.1M	33.3G	**48.0 (+28.3)**
TSM	—	—	23.9M	32.9G	45.6
	SE3D	—	26.4M	32.9G	46.7 (+1.1)
	GE3D-G	—	23.9M	32.9G	45.7 (+0.1)
	GE3D-C	—	25.2M	33.3G	47.0 (+1.4)
	S3D-G	—	25.1M	32.9G	46.8 (+1.2)
	NLN	—	31.2M	49.4G	47.2 (+1.6)
	GC	$(\frac{1}{2}, \frac{1}{2}C)$	24.2M	33.0G	48.7 (+3.1)
		$(1, C)$	25.1M	33.3G	**48.9 (+3.3)**
		$(1, C)$, loop	25.1M	33.3G	**48.9 (+3.3)**
GST	—	—	21.0M	29.2G	44.4
	GC	$(\frac{1}{2}, \frac{1}{2}C)$	21.4M	29.3G	45.5 (+1.1)
		$(1, C)$	22.3M	29.6G	45.6 (+1.2)
		$(1, C)$, loop	22.3M	29.6G	**46.7 (+2.3)**
TDN	—	—	26.1M	36.0G	52.3
	GC	$(1, C)$	27.4M	36.7G	**53.7 (+1.4)**
		$(1, C)$, loop	27.4M	36.7G	53.6 (+1.3)

- Both the single Ecal and the combined GC consistently improve the recognition performance of backbones.

Example Demonstration:



- GC can boost the recognition of activities that need global/&local contexts.

Result on Something-Something V2:

Method	Params	#Frame	FLOPs×Clips	Top-1	Top-5
TIN [36]	24.6M	16	67.0G×1	60.1	86.4
RubiksNet [5]	8.5M	8	15.8G×2	61.7	87.3
TSM+TPN [51]	—	8	33.0G×1	62.0	—
SlowFast [7]	32.9M	4+32	65.7G×6	61.9	87.0
SlowFast(R101) [7]	53.3M	8+32	106G×6	63.1	87.6
SmallBig [22]	—	16	114.0G×6	63.8	88.9
STM [19]	24.0M	16	33.3G×30	64.2	89.8
TEA [23]	—	16	70.0G×30	65.1	89.9
TEINet [27]	30.4M×2	8+16	99.0G×1	65.5	89.8
TANet [29]	25.1M×2	8+16	99.0G×6	66.0	90.1
TimeSformer-HR [2]	121.4M	16	1703G×3	62.5	—
ViViT-L [1]	352.1M	32	903G×4	65.4	89.8
MViT-B [4]	36.6M	64	455G×3	67.7	90.9
Video-Swin-B [28]	88.8M	16	321G×3	**69.6**	**92.7**
TSN [56] from [26]	23.9M	8	32.9G×1	30.0	60.5
GST* [30]	21.0M	8	29.2G×2	59.8	86.3
GST* [30]	21.0M	16	58.4G×2	61.7	87.2
GST* [30]	21.0M×2	8+16	87.6G×2	63.1	88.3
TSM [26]	23.9M	8	32.9G×2	61.2	87.1
TSM [26]	23.9M	16	65.8G×2	63.1	88.2
TSM [26]	23.9M×2	8+16	98.7G×2	64.3	89.0
TDN [45]	26.1M	8	36.0G×1	64.0	88.8
TDN [45]	26.1M	16	72.0G×1	65.3	89.5
TDN [45]	26.1M×2	8+16	108G×1	67.0	90.3
GC-GST	22.3M	8	29.6G×2	61.9	87.8
GC-GST	22.3M	16	59.1G×2	63.3	88.5
GC-GST	22.3M×2	8+16	88.7G×2	65.0	89.5
GC-TSN	25.1M	8	33.3G×2	62.4	87.9
GC-TSN	25.1M	16	66.5G×2	64.8	89.4
GC-TSN	25.1M	8+16	99.8G×2	66.3	90.3
GC-TSM	25.1M	8	33.3G×2	63.0	88.4
GC-TSM	25.1M	16	66.5G×2	64.9	89.7
GC-TSM	25.1M×2	8+16	99.8G×2	66.7	90.6
GC-TSM	25.1M×2	8+16	99.8G×6	67.5	90.9
GC-TDN	27.4M	8	36.7G×1	64.9	89.7
GC-TDN	27.4M	16	73.4G×1	65.9	90.0
GC-TDN	27.4M×2	8+16	110.1G×1	**67.8**	**91.2**

- The GC module boosts the 2D/3D video CNNs with substantial improvements on Something-Something V2 dataset.
- GC-Nets achieve either better or the best performances compared to the SOTAs.

Result on Kinetics-400:

Model	Params	#Frame	FLOPs×Clips	Top1	Top5
I3D (InceptionV1) [3]	—	64	—	72.1	90.3
Nonlocal-I3D [47]	35.3M	32	282G×10	74.9	91.6
S3D-G (InceptionV1) [50]	—	64	71.4G×30	74.7	**93.4**
TEA [23]	—	16	70G×30	76.1	92.5
TEINet [27]	30.8M	16	66G×30	76.2	92.5
TANet [29]	25.6M	16	86G×12	76.9	92.9
SmallBig [22]	—	8	57G×30	76.3	92.5
SlowFast(8×8) [7]	32.9M	8+32	65.7G×30	77.0	92.6
X3D-L [6]	6.1M	16	24.8G×30	77.5	92.9
TSN [56]	24.3M	8	32.9G×10clip	70.6	89.2
TSM [26]	24.3M	16	66.0G×10	74.7	91.4
TDN [45]	26.6M	8+16	108.0G×30	78.4	93.6
GC-TSN	25.6M	8	33.3G×10	75.2	92.1
GC-TSM	25.6M	8	33.3G×10	75.4	91.9
GC-TSM	25.6M	16	66.6G×10	76.7	92.9
GC-TSM	25.6M	16	66.6G×30	77.1	92.9
GC-TDN	27.4M	8	36.7G×30	77.3	93.2
GC-TDN	27.4M	16	73.4G×30	78.8	93.8
GC-TDN	27.4M	8+16	110.1G×30	**79.6**	94.1

- GC improves the performance of TSN, TSM and TDN by large margins on Kinetics-400 dataset, and GC-TDN with 8+16 frames achieves the highest top-1 accuracy of 79.6% over all the competing methods.