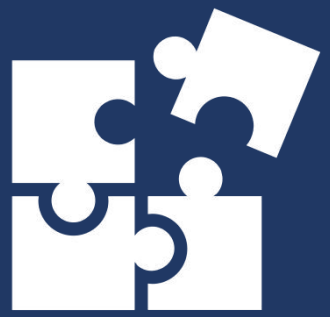


VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, Yu Qiao

Nanjing University, Shanghai AI Lab, Shenzhen Institute of Advanced Technology

CVPR 2023

PRESENTER: LILANG LIN

2023/05/14

Outline

1 Authors

2 **Background**

3 Method

4 Experiments

5 Discussion

Background

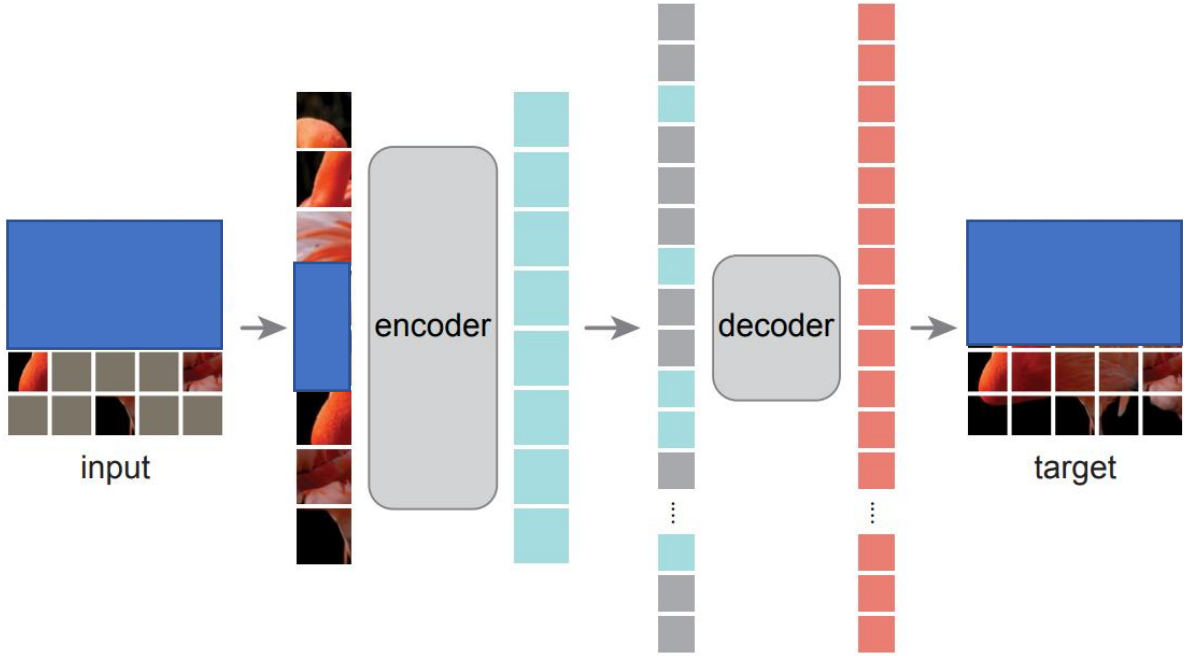
MAE (CVPR 2022)

Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

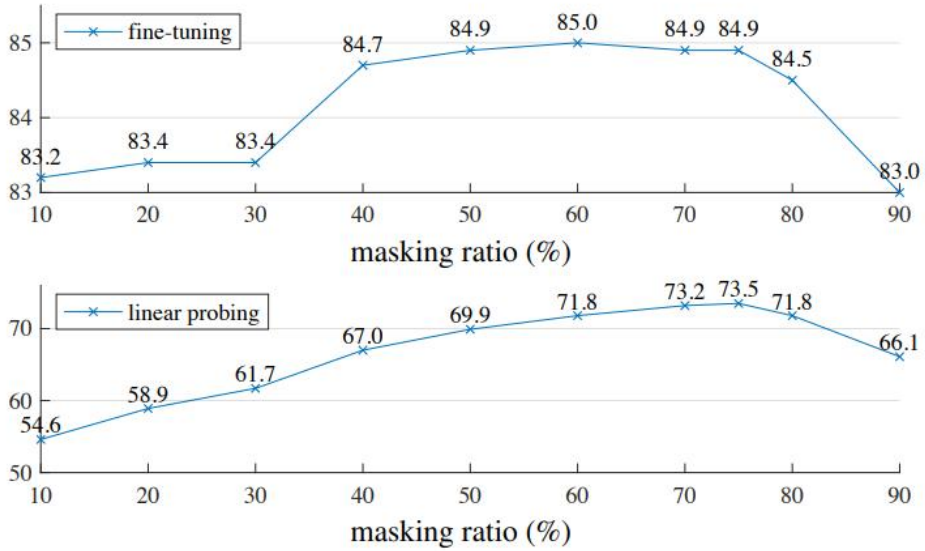
*equal technical contribution †project lead

Facebook AI Research (FAIR)



Background

MAE (CVPR 2022)

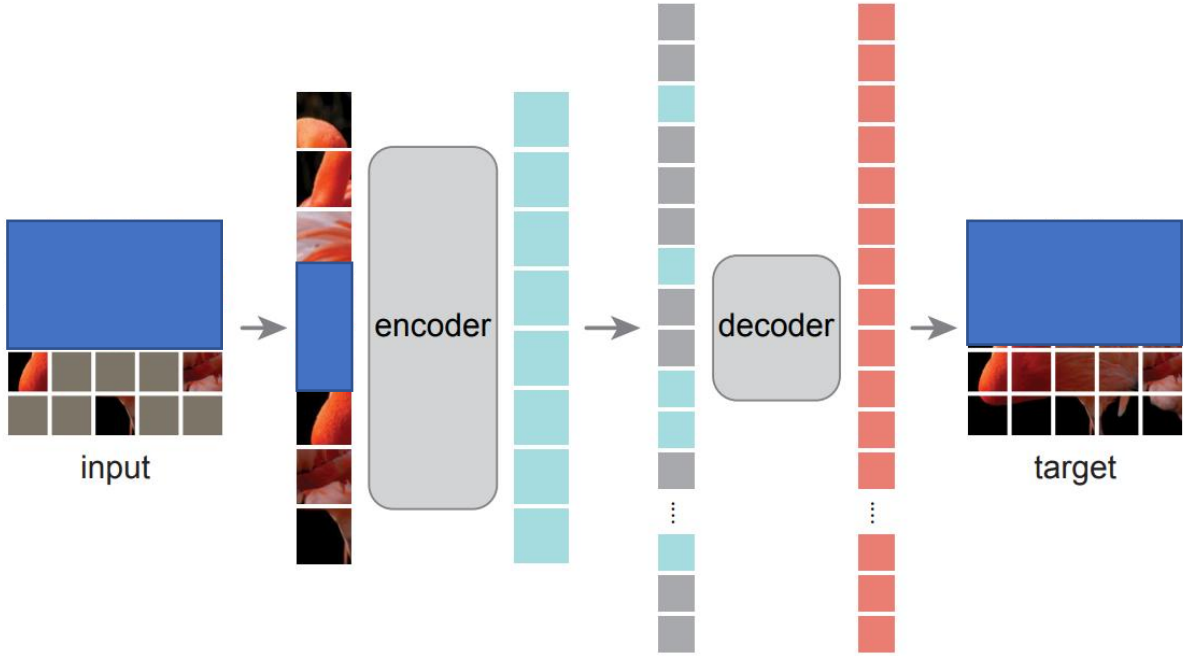


Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)



Background

MAE (CVPR 2022)

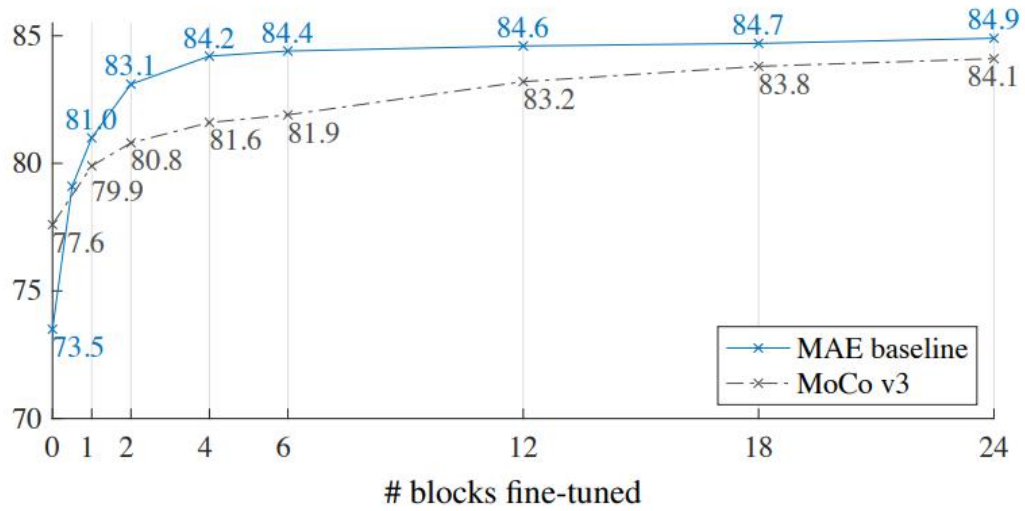
Masked Autoencoders Are Scalable Vision Learners

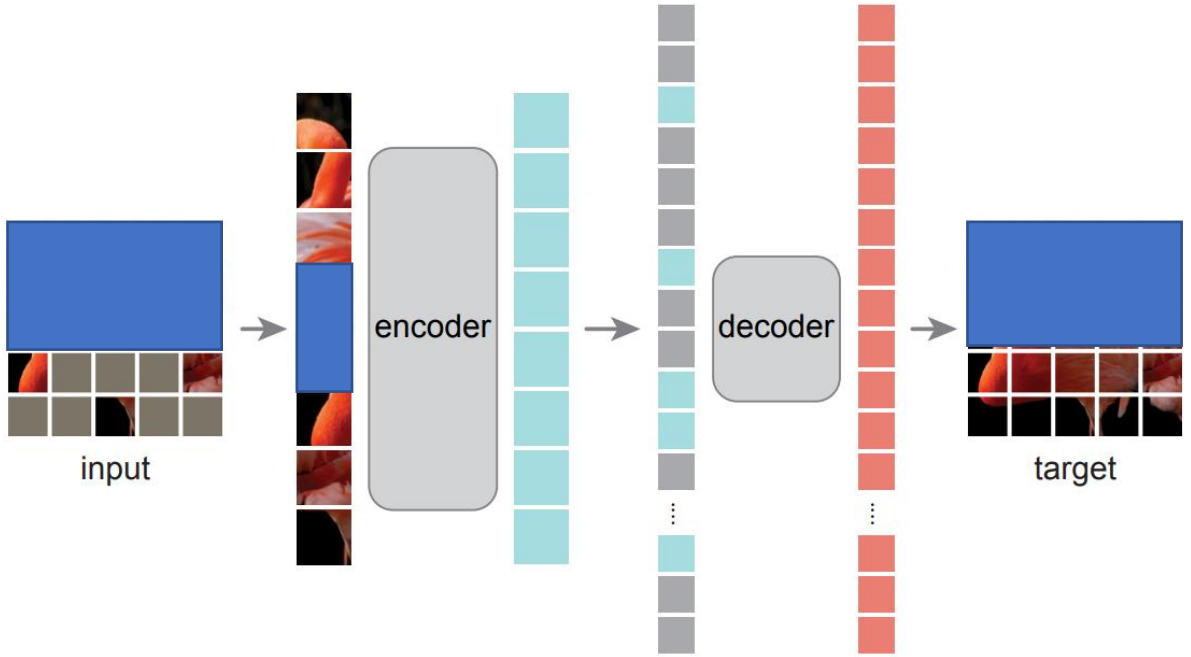
Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**





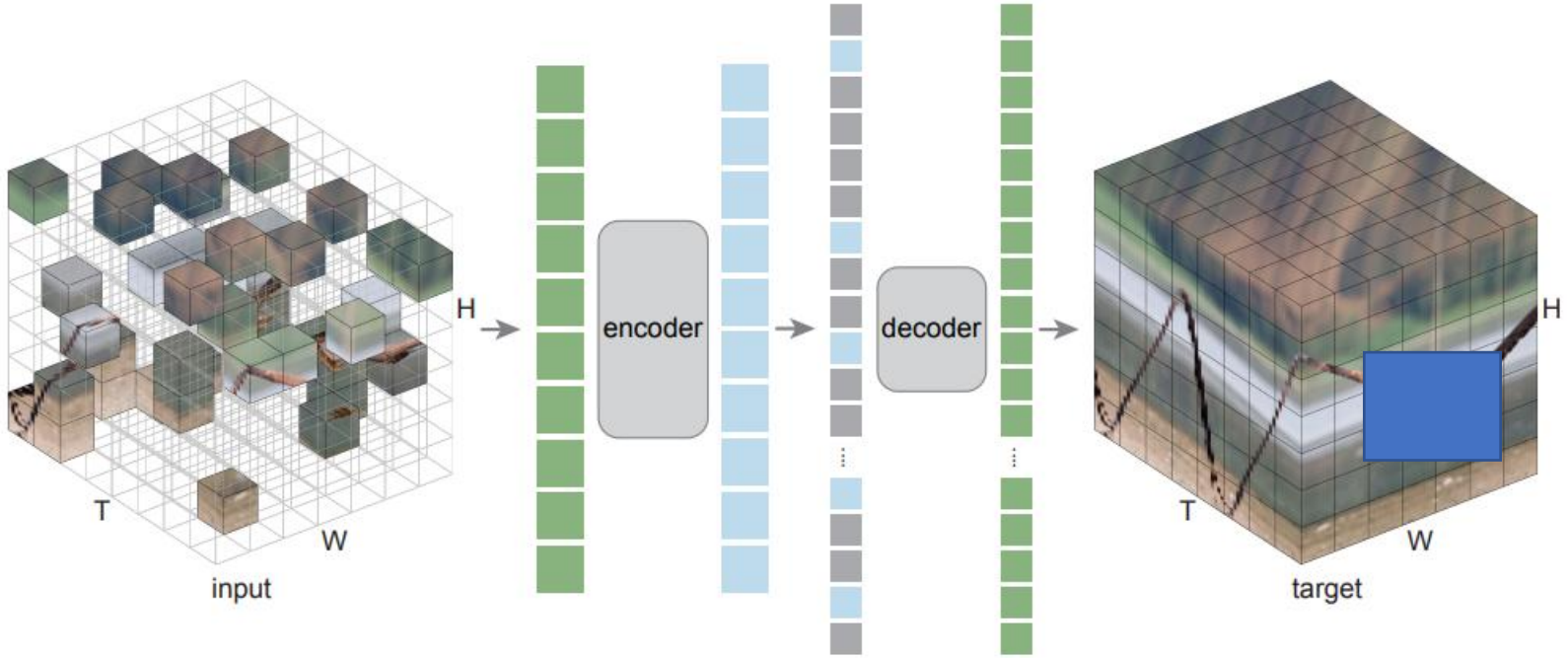
Background

Video MAE

Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer* Haoqi Fan* Yanghao Li Kaiming He
Meta AI, FAIR

https://github.com/facebookresearch/mae_st



Background

Video MAE

Masked Autoencoders As Spatiotemporal Learners

Christoph Feichtenhofer* **Haoqi Fan*** **Yanghao Li** **Kaiming He**
Meta AI, FAIR

https://github.com/facebookresearch/mae_st

pre-train set	# pre-train data	pre-train method	K400	AVA	SSv2
-	-	none (from scratch)	71.4	-	-
IN1K	1.28M	supervised	78.6	17.3	50.2
IN1K	1.28M	MAE	82.3	26.3	65.6
K400	240k	supervised	-	21.6	55.7
K400	240k	MAE	84.8	31.1	72.1
K600	387k	MAE	**84.9**	32.5	73.0
K700	537k	MAE	n/a†	33.1	**73.6**
IG-uncurated	1M	MAE	84.4	**34.2**	**73.6**

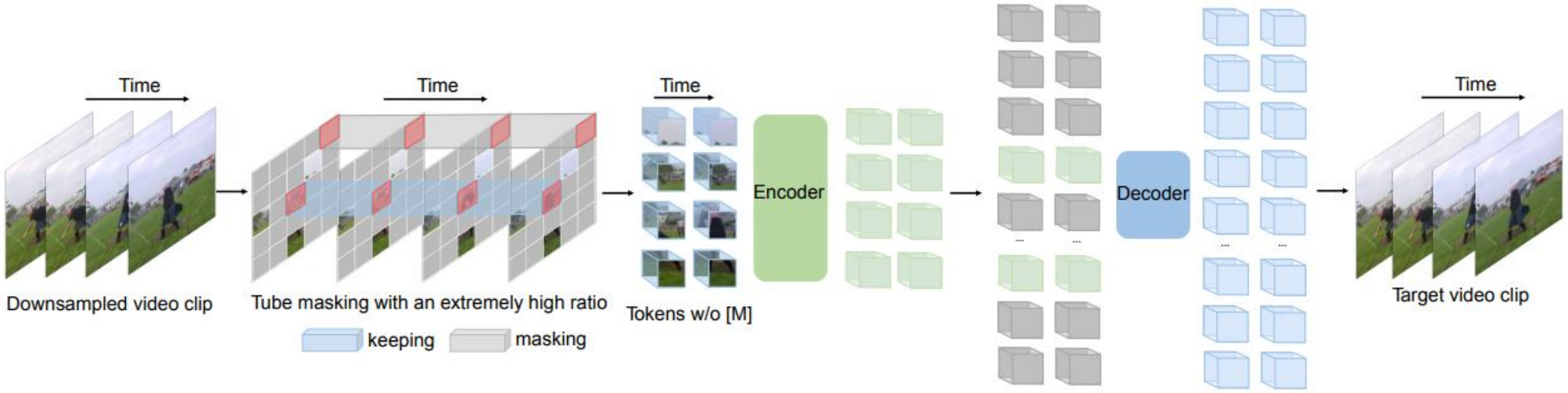
MAE		ViT-B	16×224^2	81.3	94.9	$180 \times 3 \times 7$	87
MAE		ViT-L	16×224^2	84.8	96.2	$598 \times 3 \times 7$	304
MAE		ViT-H	16×224^2	85.1	96.6	$1193 \times 3 \times 7$	632
MAE		ViT-L	40×312^2	85.8	96.9	$4757 \times 3 \times 7$	304
MAE		ViT-H	32×312^2	86.0	97.0	$6382 \times 3 \times 7$	632

Background

VideoMAE (NeurIPS 2022)

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn



Background

VideoMAE (NeurIPS 2022)

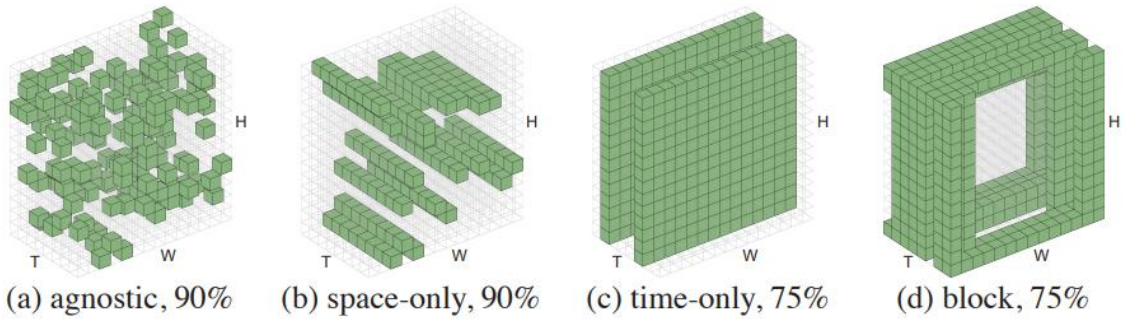
VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn

dataset	training data	*from scratch*	MoCo v3	VideoMAE
K400	240k	68.8	74.2	**80.0**
Sth-Sth V2	169k	32.6	54.2	**69.6**
UCF101	9.5k	51.4	81.7	**91.3**
HMDB51	3.5k	18.0	39.2	**62.6**

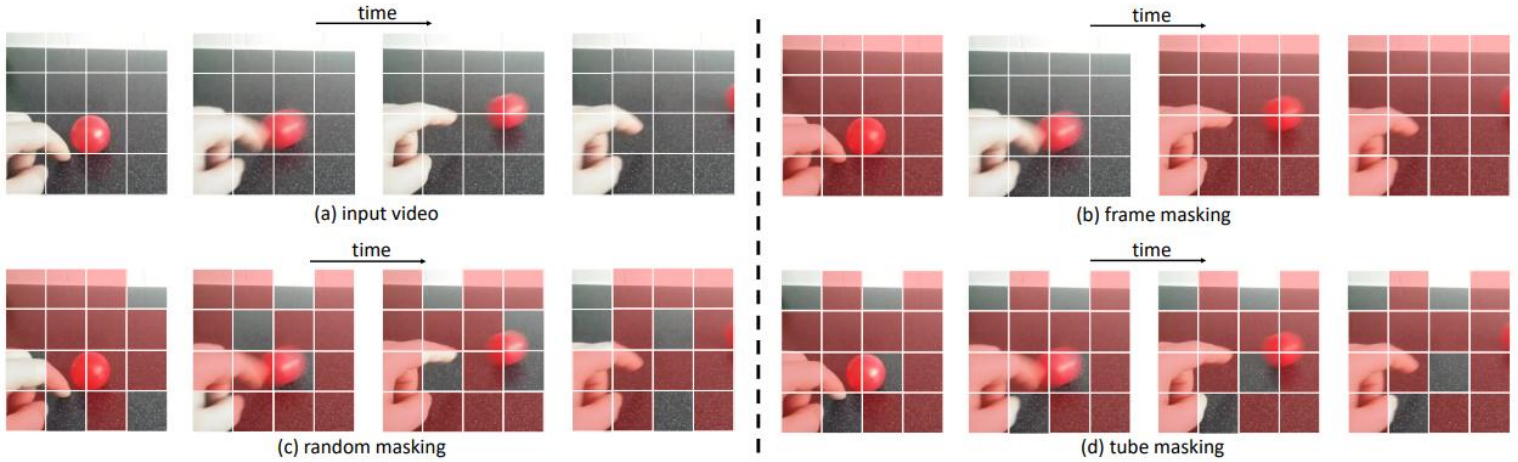
Background

VideoMAE vs VideoMAE



(a) agnostic, 90% (b) space-only, 90% (c) time-only, 75% (d) block, 75%

case	ratio	acc.
agnostic	90	**84.4**
space-only	90	83.5
time-only	75	79.1
block	75	83.2



(a) input video (b) frame masking (c) random masking (d) tube masking

case	ratio	SSV2	K400
tube	75	68.0	79.8
tube	90	**69.6**	**80.0**
random	90	68.3	79.5
frame	87.5*	61.5	76.5

Background

AdaMAE (CVPR 2023)

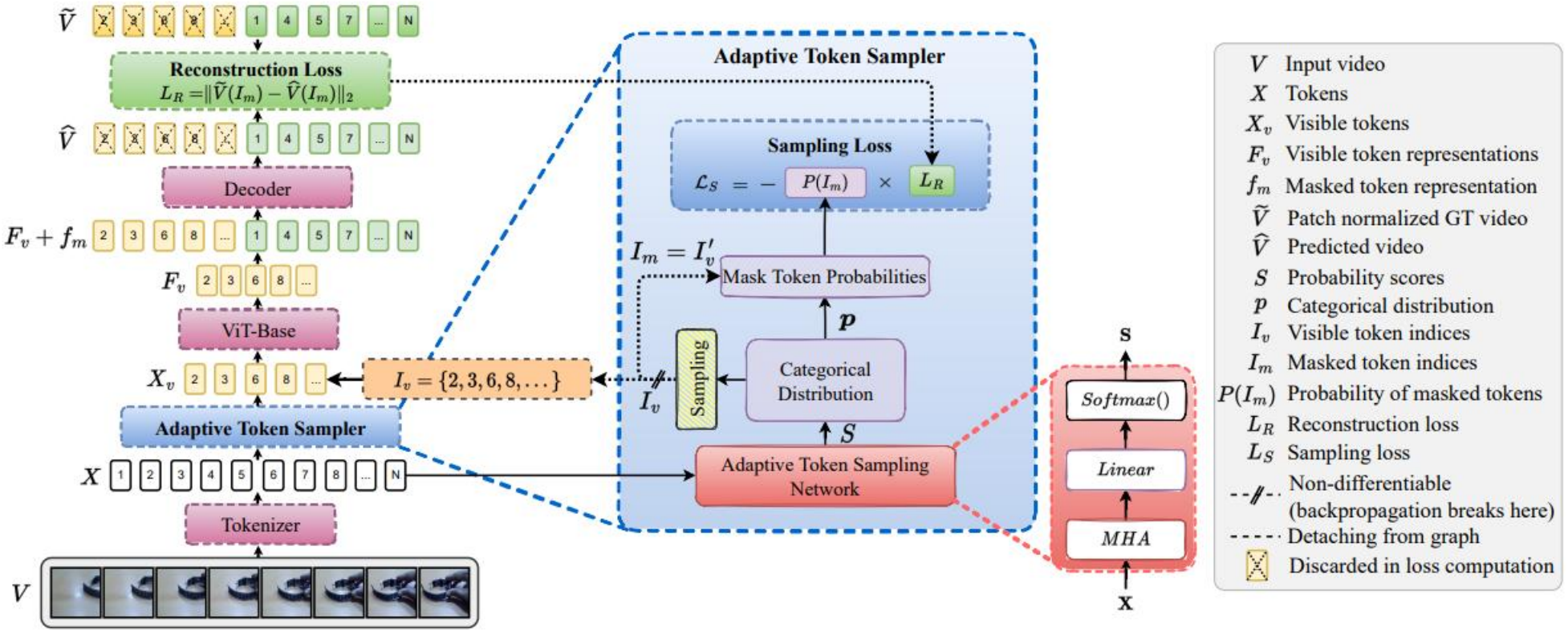
AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders

Wele Gedara Chaminda Bandara[1], Naman Patel[2], Ali Gholami[2], Mehdi Nikkhah[2], Motilal Agrawal[2], and Vishal M. Patel[1]

[1]Johns Hopkins University, Baltimore, USA [2]Zippin, California, USA

wbandar1@jhu.edu, {naman, gholami, mehdi, moti}@getzippin.com, vpatel36@jhu.edu

github.com/wgcban/adamae



Background

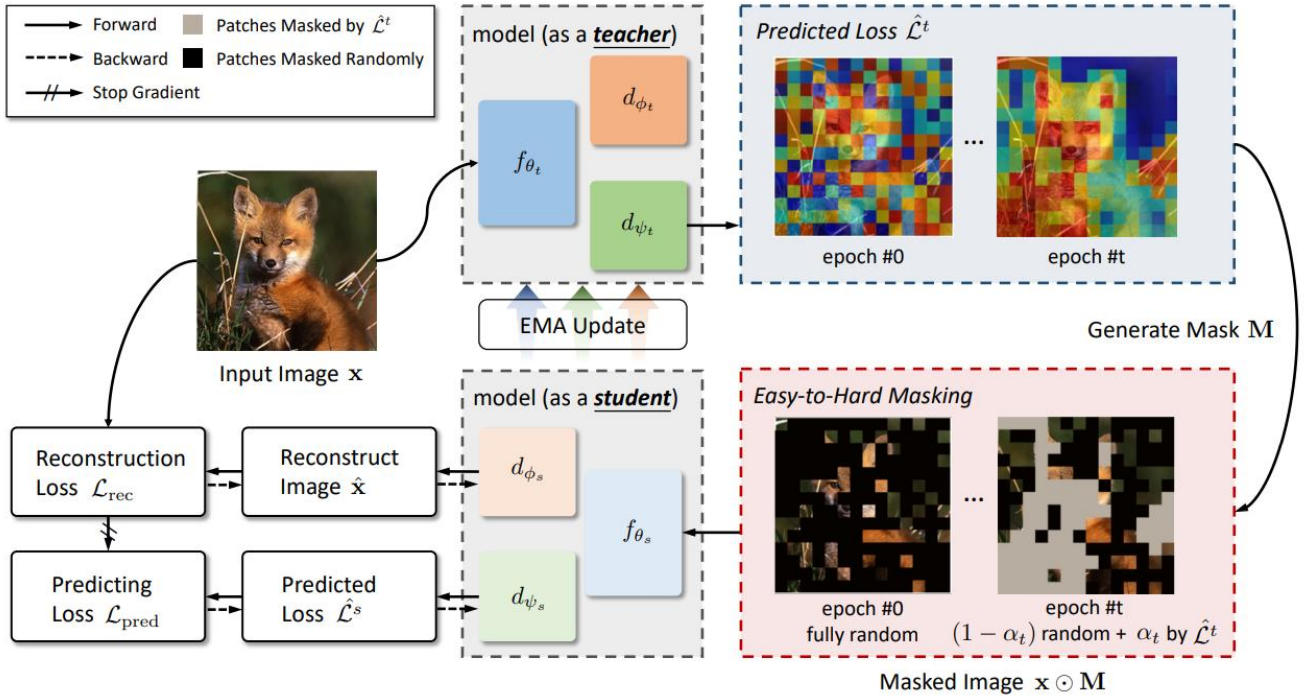
HPM (CVPR 2023)

Hard Patches Mining for Masked Image Modeling

Haochen Wang[1,3] Kaiyou Song[2] Junsong Fan[1,4] Yuxi Wang[1,4] Jin Xie[2] Zhaoxiang Zhang[1,3,4]

[1]Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Megvii Technology [3]University of Chinese Academy of Sciences
[4]Centre for Artificial Intelligence and Robotics,
Hong Kong Institute of Science & Innovation, Chinese Academy of Science

{wanghaochen2022, junsong.fan, zhaoxiang.zhang}@ia.ac.cn
{songkaiyou, xiejin}@megvii.com yuxiwang93@gmail.com



Outline

1 Authors

2 Background

3 **Method**

4 Experiments

5 Discussion

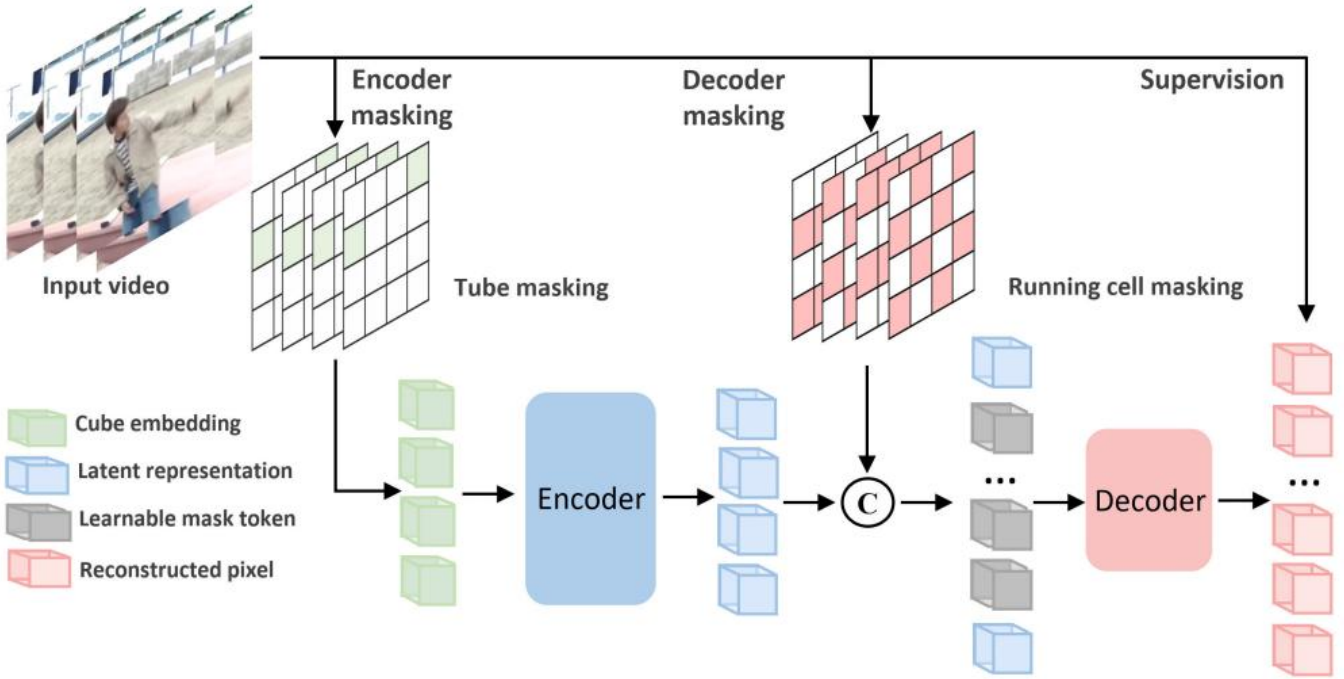
Method

Dual Masking for VideoMAE

Masking Decoder

$$\mathbf{Z}^c = \mathbf{Z} \cup \{\mathbf{M}_i\}_{i \in \mathbb{M}_d}$$

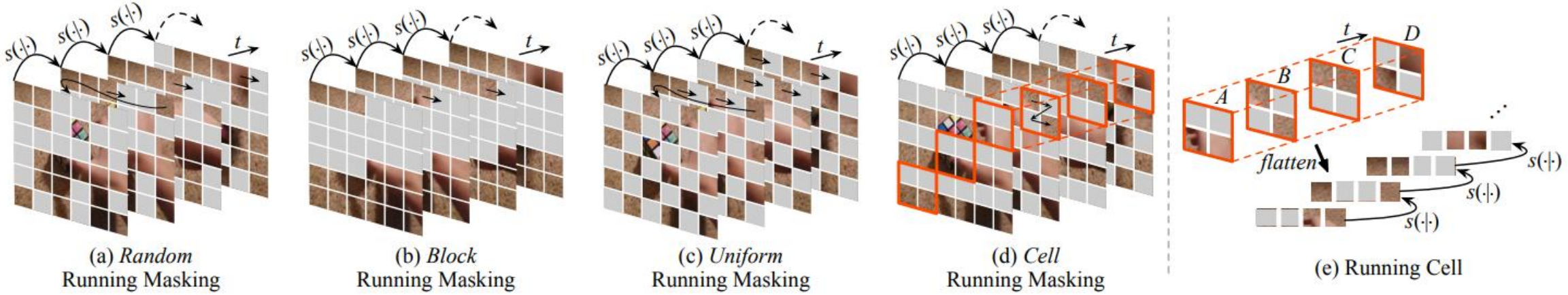
$$\ell = \frac{1}{(1-\rho^d)N} \sum_{i \in \mathbb{M}_d \cap \mathbb{M}_e} |\mathbf{I}_i - \hat{\mathbf{I}}_i|^2$$



Method

Dual Masking for VideoMAE

Cell Running Masking

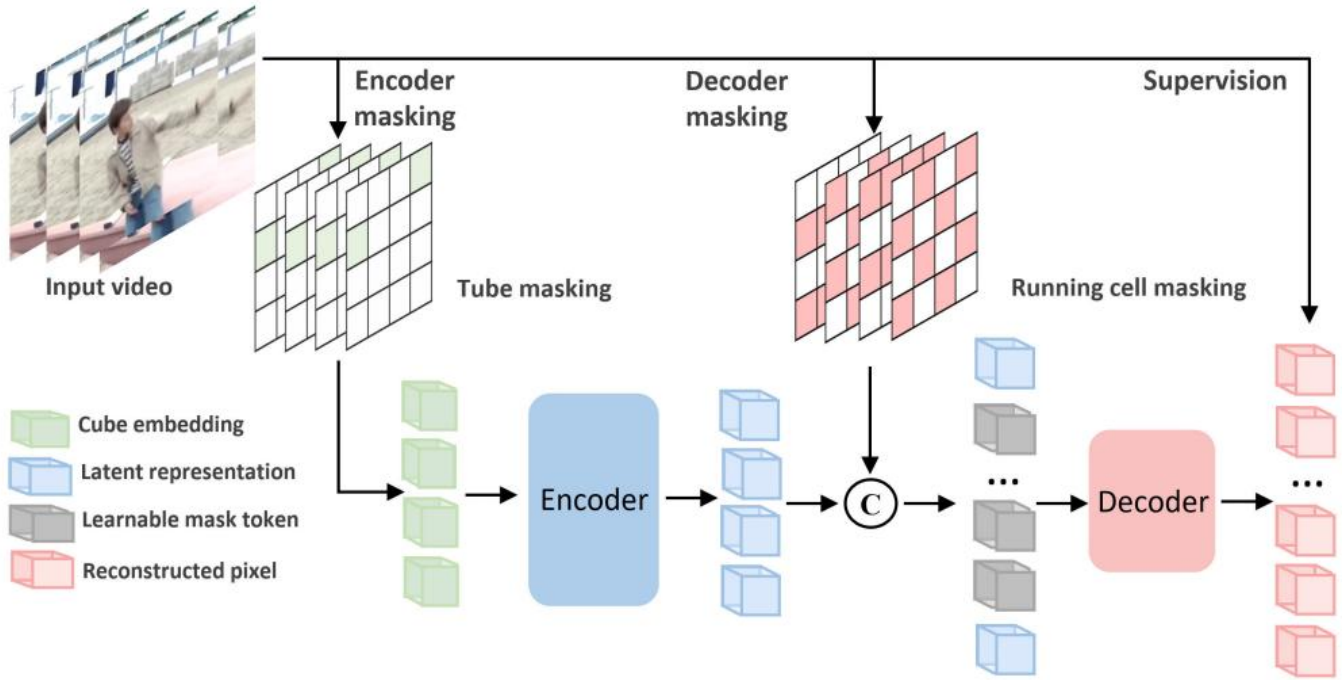


(a) *Random* Running Masking (b) *Block* Running Masking (c) *Uniform* Running Masking (d) *Cell* Running Masking (e) Running Cell

$$M_t = s(M_{t-1}|M_{t-2}, \dots M_1),$$

Method

- Scaling VideoMAE
  - Model Scaling
  - Data Scaling
  - Progressive Training



Outline

1 Authors

2 Background

3 Method

4 **Experiments**

5 Discussion

Experiments

Action Classification

method	pre-train data	data size	epoch	ViT-B	ViT-L	ViT-H	ViT-g
MAE-ST [28]	Kinetics400	0.24M	1600	81.3	84.8	85.1	-
MAE-ST [28]	IG-uncurated	1M	1600	-	84.4	-	-
VideoMAE V1 [90]	Kinetics400	0.24M	1600	**81.5**	85.2	86.6	-
VideoMAE V2	UnlabeledHybrid	1.35M	1200	**81.5** (77.0)	**85.4** (81.3)	**86.9** (83.2)	**87.2** (83.9)
Δ*Acc.* with V1	-	-	-	*+ 0%*	*+ 0.2%*	*+ 0.3%*	-

Table 3. **Results on the Kinetics-400 dataset**. We scale the pre-training of VideoMAE V2 to billion-level ViT-g model with million-level data size. We report the fine-tuning accuracy of multiple view fusion (5×3) and single view results in the bracket. All models are pre-trained and fine-tuned at the input of 16×224×224 and sampling stride $\tau = 4$.

method	pre-train data	data size	epoch	ViT-B	ViT-L	ViT-H	ViT-g
MAE-ST [28]	Kinetics400	0.24M	1600	-	72.1	74.1	-
MAE-ST [28]	Kinetics700	0.55M	1600	-	73.6	75.5	-
VideoMAE V1 [90]	Something-Something V2	0.17M	2400	70.8	74.3	74.8	-
VideoMAE V2	UnlabeledHybrid	1.35M	1200	**71.2** (69.5)	**75.7** (74.00)	**76.8** (75.5)	**77.0** (75.7)
Δ*Acc.* with V1	-	-	-	*+ 0.4%*	*+ 1.4%*	*+ 2.0%*	-

Table 4. **Results on the Something-Something V2 dataset**. We scale the pre-training of VideoMAE V2 to billion-level ViT-g model with million-level data size. We report the fine-tuning accuracy of multiple view fusion (2×3) and single view results in the brackets. All models are pre-trained at input of 16×224×224 and sampling stride $\tau = 2$. Fine-tuning is on the same size as TSN [99] sampling.

Experiments

Action Detection

(e) AVA

Method	Long Feature	mAP
SlowFast [29]	✗	29.0
TubeR [120]	✓	33.4
MaskFeat [104]	✗	38.8
MAE-ST [28]	✗	39.0
VideoMAE [90]	✗	39.5
VideoMAE V2	✗	**42.6**

(f) AVA Kinetics

Method	Ensembled	mAP
AIA++ [106]	✓	29.0
MSF [121]	✓	33.4
ACAR [72]	✓	40.5
VideoMAE V2	✗	**43.9**

(g) THUMOS14

Method	Optical Flow	mAP
RTD-Net [87]	✓	43.6
DaoTAD [94]	✗	50.0
AFSD [57]	✓	52.0
DCAN [14]	✓	52.3
TadTR [60]	✓	54.2
TALLFormer [19]	✗	59.2
BasicTAD [113]	✗	59.6
ActionFormer [117]	✓	66.8
VideoMAE V2	✗	**69.6**

(h) FineAction

Method	Optical Flow	mAP
BMN [59]	✓	9.25
G-TAD [110]	✓	9.06
BasicTAD [113]	✗	12.2
ActionFormer [117]	✗	13.2
VideoMAE V2	✗	**18.2**

Experiments

Decoder Masking Strategies

Decoder Masking	ρ^d	Top-1	FLOPs
None	0%	**70.28**	35.48G
Frame	50%	69.76	25.87G
Random	50%	64.87	25.87G
Running cell [1]	50%	66.74	25.87G
Running cell [2]	25%	70.22	31.63G
Running cell [2]	50%	70.15	25.87G
Running cell [2]	75%	70.01	21.06G

Experiments

Dual Masking & Encoder-Only Masking

Masking	Backbone	pre-training dataset	FLOPs	Mems	Time	Speedup	Top-1
Encoder masking	ViT-B	Something-Something V2	35.48G	631M	28.4h	-	70.28
Dual masking	ViT-B	Something-Something V2	25.87G	328M	15.9h	**1.79**×	70.15
Encoder masking	ViT-g	UnlabeledHybrid	263.93G	1753M	356h[1]	-	-
Dual masking	ViT-g	UnlabeledHybrid	241.61G	1050M	241h	**1.48**×	77.00

Experiments

Progressive Pre-Training

method	extra supervision	ViT-H	ViT-g
MAE-ST [28]	K600	86.8	-
VideoMAE V1 [90]	K710	88.1 (84.6)	-
VideoMAE V2	-	86.9 (83.2)	87.2 (83.9)
VideoMAE V2	K710	**88.6** (85.0)	**88.5** (85.6)
Δ*Acc.* with V1	K710	***+ 0.5%***	-

Outline

1 Authors

2 Background

3 Method

4 Experiments

5 **Discussion**

Discussion

- Masked Modeling

Thanks!