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

Zhan Tong, Yibing Song, Jue Wang, Limin Wang
NeurIPS 2022

Presented by Liujie Zheng, Junjie Zhao

Motivation

Effective video representation learning improves downstream tasks

- e.g. action detection

Challenges for video understanding

- temporal redundancy and correlation
- higher computational consumption for video

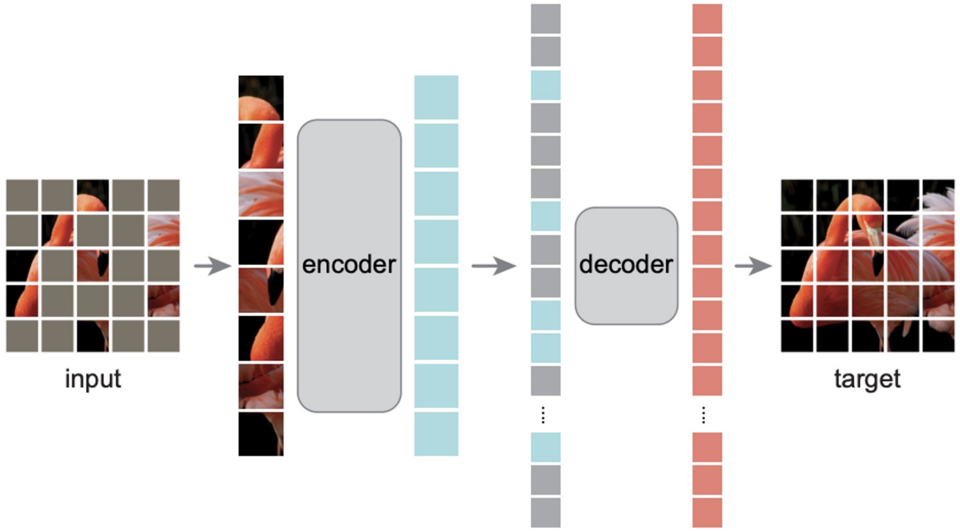
Challenges for training video transformer

- need extra large-scale image/video data
- heavily depend on pre-trained models (e.g. ImageNet-1K)

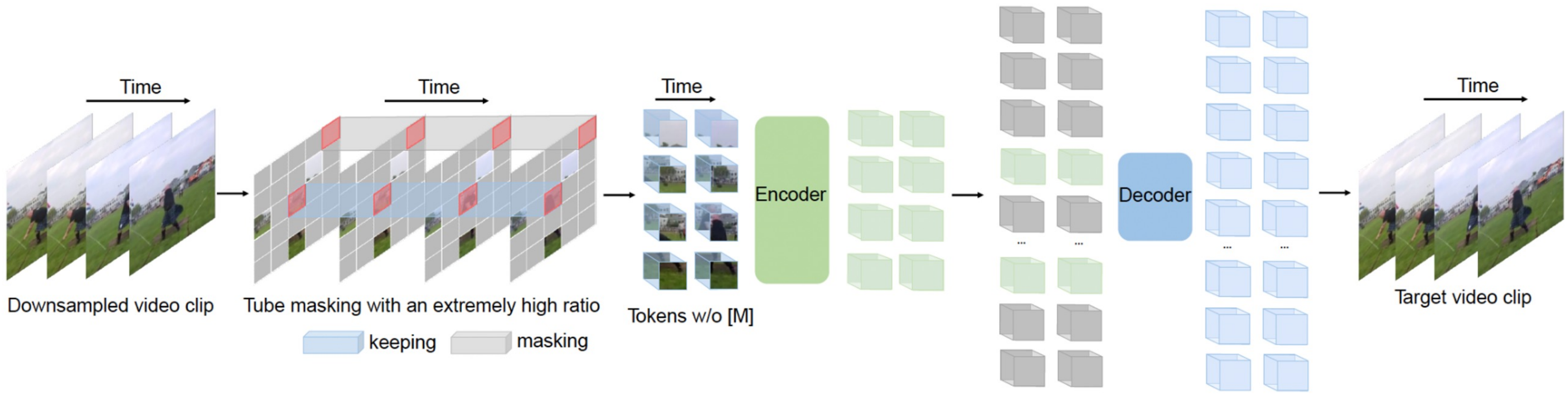
How to efficiently train a vanilla ViT on the video dataset itself
without using any pre-trained model or extra data?

Inspiration: ImageMAE

- Mask random patches of the input image and reconstruct the missing pixels
- An asymmetric encoder-decoder architecture



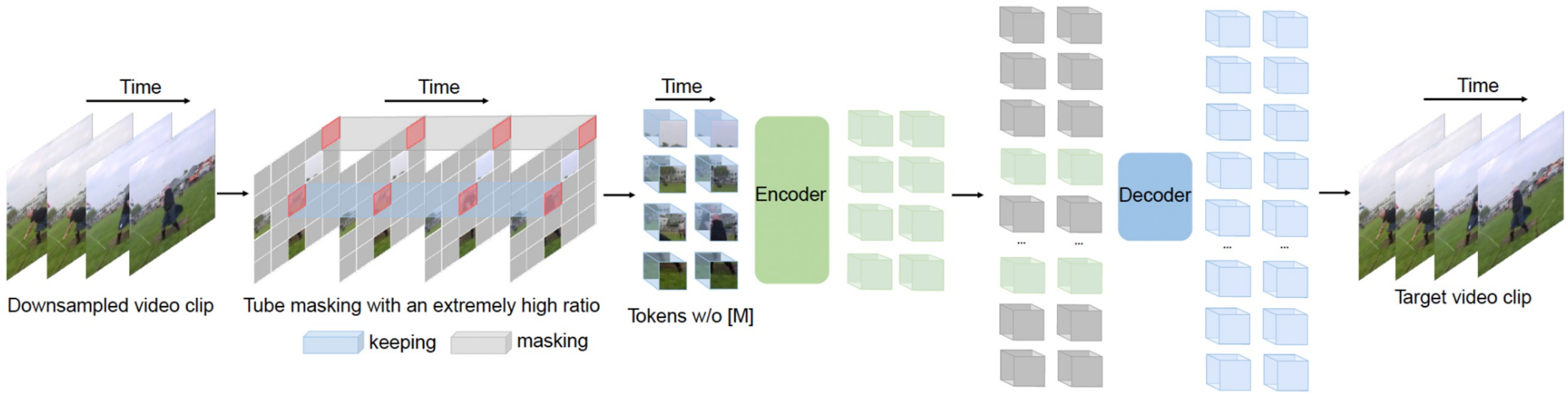
VideoMAE



Self-supervised pre-training with masked autoencoder

- a simple but effective masking and reconstruction proxy task
- an efficient pre-training process with only unmasked tokens into the encoder.

VideoMAE



A new masking strategy:

- tube masking with an extremely high ratio (90%-95%)
- making video reconstruction a more challenging self-supervision task

New Masking Strategy

Temporal redundancy: the semantics vary slowly in the temporal dimension

- less efficient to keep the original temporal frame rate
- greatly dilutes motion representations, making the task of reconstructing missing pixels not difficult
- Solution: high mask ratio (90%-95%)

Temporal correlation: inherent correspondence between adjacent frames

- we can reconstruct the masked patches by finding the spatiotemporal corresponding unmasked patches in the adjacent frames
- Solution: tube mask (the masking map is the same for all frames)

VideoMAE Architecture

Stage	Vision Transformer (Base)	Output Sizes
data	stride $4\times1\times1$ on *K400* stride $2\times1\times1$ on *SSV2*	$3\times16\times224\times224$
cube	$2\times16\times16$, 768 stride $2\times16\times16$	$768\times\mathbf{8}\times196$
mask	tube mask *mask ratio* = ρ	$768\times\mathbf{8}\times[196\times(1\text{-}\rho)]$
encoder	$\begin{bmatrix}\text{MHA}(768)\\ \text{MLP}(3072)\end{bmatrix}\times12$	$768\times\mathbf{8}\times[196\times(1\text{-}\rho)]$
projector	MLP(384) & *concat learnable tokens*	$384\times\mathbf{8}\times196$
decoder	$\begin{bmatrix}\text{MHA}(384)\\ \text{MLP}(1536)\end{bmatrix}\times4$	$384\times\mathbf{8}\times196$
projector	MLP(1536)	$1536\times\mathbf{8}\times196$
reshape	*from* 1536 *to* $3\times2\times16\times16$	$3\times16\times224\times224$

Output sizes are denoted by {**C**×**T**×**S**}

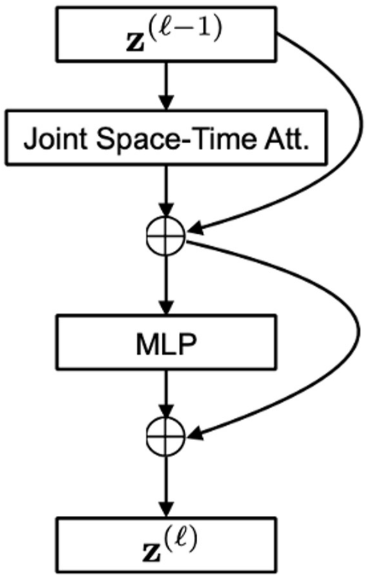
- Uses the Vit-Base for example
- Tested with Vit-Large and Vit-Huge

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Details of Vision Transformer model variants

VideoMAE Architecture

- Uses the vanilla ViT backbone
- High proportion of masking ratio
- Joint space-time attention



Experiments

Evaluated on five video datasets:

- Kinetics-400 (240k training videos)
- Something-Something V2 (169k training videos)
- UCF101 (9.5k training videos)
    - Action recognition data set of realistic action videos, collected from YouTube, having 101 action categories
- HMDB51 (3.5k training videos)
    - Human motion recognition dataset with 51 action categories
- AVA
    - A dataset for spatiotemporal localization of human actions (Transfer learning for downstream action detection tasks)

Ablation Study

blocks	SSV2	K400	GPU mem.
1	68.5	79.0	7.9G
2	69.2	79.2	10.2G
4	**69.6**	**80.0**	14.7G
8	69.3	79.7	23.7G

Decoder Depth Choice

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

Mask sampling

case	SSV2	K400
from scratch	32.6	68.8
ImageNet-21k sup.	61.8	78.9
IN-21k+K400 sup.	65.2	-
VideoMAE	**69.6**	**80.0**

Pre-training strategy

dataset	method	SSV2	K400
IN-1K	ImageMAE	64.8	78.7
K400	VideoMAE	68.5	**80.0**
SSV2	VideoMAE	**69.6**	79.6

Pre-training dataset

Results and Analysis

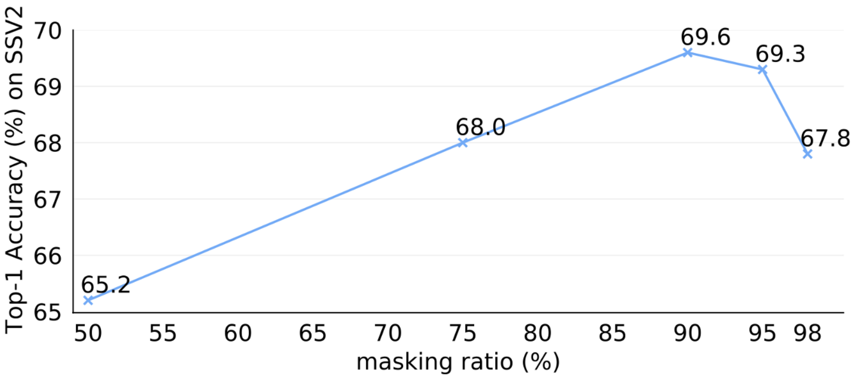
- Data-efficient learner
- VideoMAE still obtain a satisfying accuracy on small dataset like HMDB51

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**

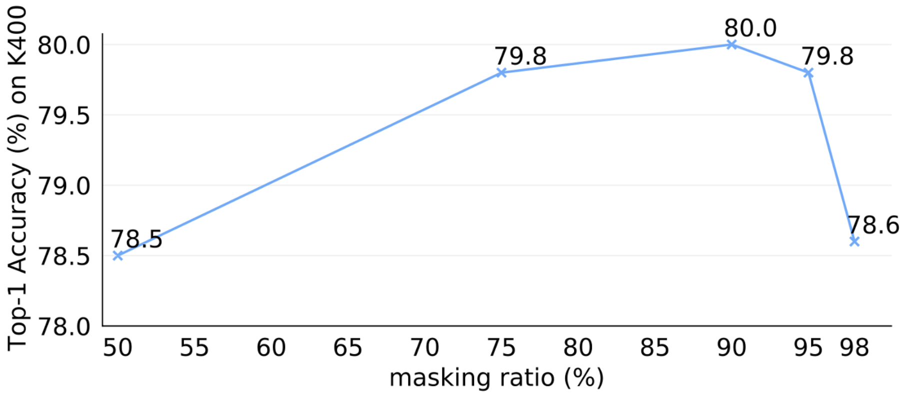
Performance on video datasets of different scales

Results and Analysis

Effectiveness of high masking ratio



Top-1 Accuracy on SSV2



Top-1 Accuracy on Kinetics-400

Results and Analysis

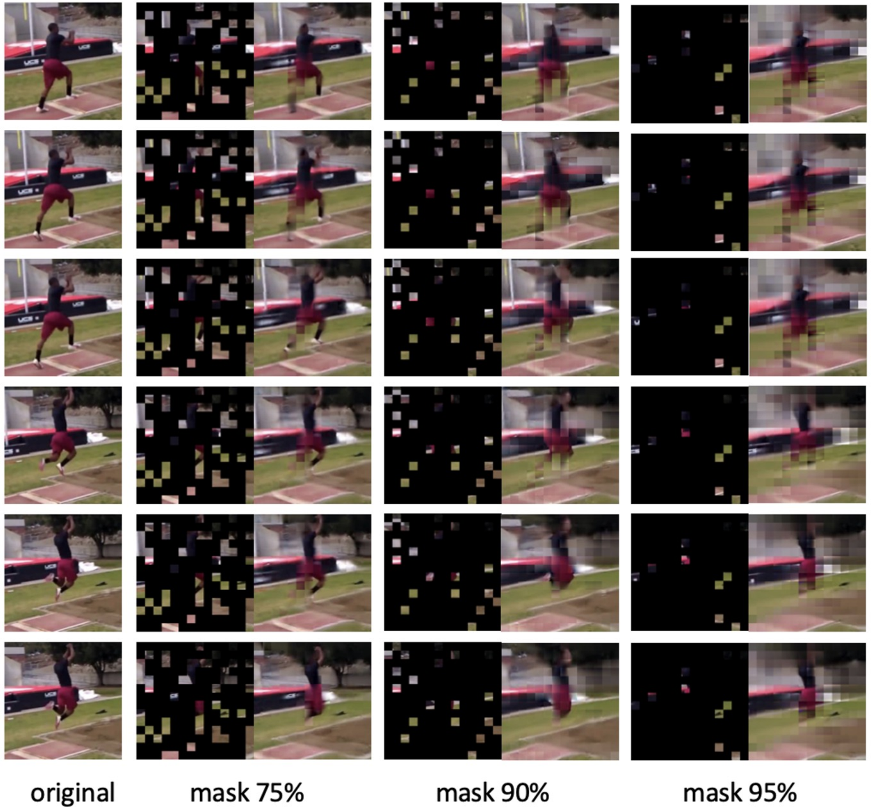
Transfer learning

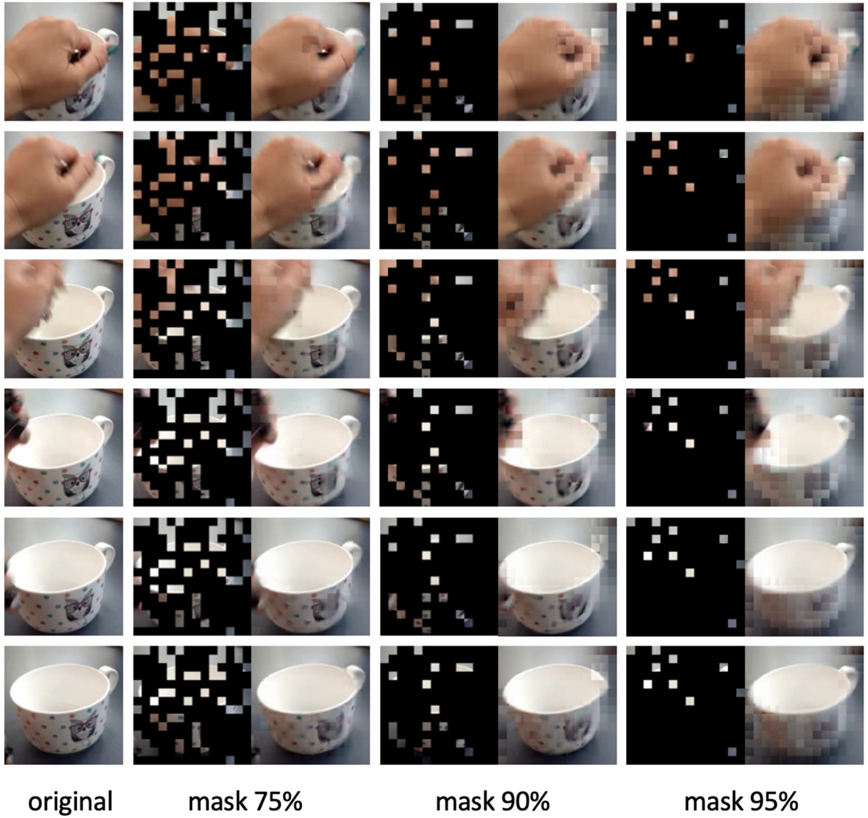
method	K400 → SSV2	K400 → UCF	K400 → HMDB
MoCo v3	62.4	93.2	67.9
VideoMAE	**68.5**	**96.1**	**73.3**

Comparisons with the feature transferability on smaller datasets

Method	Backbone	Pre-train Dataset	Extra Labels	$T \times \tau$	GFLOPs	Param	mAP
supervised [23]	SlowFast-R101	Kinetics-400	✓	8×8	138	53	23.8
CVRL [54]	SlowOnly-R50	Kinetics-400	✗	32×2	42	32	16.3
ρBYOL$_{\rho=3}$ [24]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	23.4
ρMoCo$_{\rho=3}$ [24]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	20.3
MaskFeat↑312 [80]	MViT-L	Kinetics-400	✓	40×3	2828	218	37.5
MaskFeat↑312 [80]	MViT-L	Kinetics-600	✓	40×3	2828	218	38.8
VideoMAE	ViT-S	Kinetics-400	✗	16×4	57	22	22.5
VideoMAE	ViT-S	Kinetics-400	✓	16×4	57	22	28.4
VideoMAE	ViT-B	Kinetics-400	✗	16×4	180	87	26.7
VideoMAE	ViT-B	Kinetics-400	✓	16×4	180	87	31.8
VideoMAE	ViT-L	Kinetics-400	✗	16×4	597	305	34.3
VideoMAE	ViT-L	Kinetics-400	✓	16×4	597	305	37.0
VideoMAE	ViT-H	Kinetics-400	✗	16×4	1192	633	**36.5**
VideoMAE	ViT-H	Kinetics-400	✓	16×4	1192	633	**39.5**
VideoMAE	ViT-L	Kinetics-700	✗	16×4	597	305	**36.1**
VideoMAE	ViT-L	Kinetics-700	✓	16×4	597	305	**39.3**

Visual Results





Results and Analysis

Comparison with the state-of-the-art methods on **Kinetics-400**

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [78]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [41]	ResNet152		✓	16	242×4×3	59	79.3	94.1
TDN_{En} [75]	ResNet101		✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L		✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [51]	ViT-L		✓	32	1185×10×3	382	80.2	94.8
Video Swin [39]	Swin-L		✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [77]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
MaskFeat↑352 [80]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [69]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [23]	R101+NL		✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [22]	MViTv1-B		✗	32	170×5×1	37	80.2	94.4
MaskFeat [80]	MViT-L		✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B		✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L		✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H		✗	16	1192×5×3	633	**86.6**	**97.1**
VideoMAE↑320	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
VideoMAE↑320	ViT-H		✗	32	7397×4×3	633	**87.4**	**97.6**

Results and Analysis

Comparison with the state-of-the-art methods on **Something-Something V2**

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
TEINet_{En} [40]	$\text{ResNet50}_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50	66.5	N/A
TANet_{En} [41]	$\text{ResNet50}_{\times 2}$		✓	8+16	99×2×3	51	66.0	90.1
TDN_{En} [75]	$\text{ResNet101}_{\times 2}$		✓	8+16	198×1×3	88	69.6	92.2
SlowFast [23]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53	63.1	87.6
MViTv1 [22]	MViTv1-B		✓	64	455×1×3	37	67.7	90.9
TimeSformer [6]	ViT-B	ImageNet-21K	✓	8	196×1×3	121	59.5	N/A
TimeSformer [6]	ViT-L		✓	64	5549×1×3	430	62.4	N/A
ViViT FE [3]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
Motionformer [51]	ViT-B		✓	16	370×1×3	109	66.5	90.1
Motionformer [51]	ViT-L		✓	32	1185×1×3	382	68.1	91.2
Video Swin [39]	Swin-B		✓	32	321×1×3	88	69.6	92.7
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	68.1	N/A
BEVT [77]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88	70.6	N/A
$\text{MaskFeat}_{\uparrow 312}$ [80]	MViT-L	Kinetics-600	✓	40	2828×1×3	218	75.0	95.0
VideoMAE	ViT-B	Kinetics-400	✗	16	180×2×3	87	69.7	92.3
VideoMAE	ViT-L	Kinetics-400	✗	16	597×2×3	305	74.0	94.6
VideoMAE	ViT-S	*no external data*	✗	16	57×2×3	22	66.8	90.3
VideoMAE	ViT-B		✗	16	180×2×3	87	70.8	92.4
VideoMAE	ViT-L		✗	16	597×2×3	305	74.3	94.6
VideoMAE	ViT-L		✗	32	1436×1×3	305	**75.4**	**95.2**

Thank you!