

MotionMAE: Self-supervised Video Representation Learning with Motion-Aware Masked Autoencoders

Haosen Yang¹

haosen.yang.6@gmail.com

Deng Huang⁵

im.huangdeng@gmail.com

Bin Wen²

wenbin.1994@bytedance.com

Giannan Wu²

wjn922@connect.hku.hk

Hongxun Yao⁴

h.yao@hit.edu.cn

Yi Jiang²

jiangyi0425@gmail.com

Xiatian Zhu¹

xiatian.zhu@surrey.ac.uk

Zehuan Yuan²

yuanzehuan@bytedance.com

¹ University of Surrey

Surrey, UK

² Bytedance, Inc

Beijing, China

³ The University of Hong Kong

Hong Kong, China

⁴ Harbin Institute of Technology

Harbin, China

⁵ South China University of Technology

Guangzhou, China

A Additional Implementation Details

We pretrain MotionMAE on Something-Something V2, Kinetics-400 and UCF101 using the hyper-parameters as summarized in Table 1(a). The dataset specific hyper-parameters are given in the individual columns, with the others shared across datasets. These settings apply to ViT-B and ViT-L, unless specified otherwise. We use the same hyper-parameters for pretraining. Table 1(b) summarizes our fine-tuning settings.

We conduct the experiments with 32 A100 GPUs for both pretraining and finetuning on Something-Something V2 and Kinetics-400 datasets. The experiments on smaller UCF101 are trained on 8 V100 GPUs.

B Additional Experimental Results

UCF101 Except large action datasets as above, we further evaluate on the small UCF101 dataset. As dataset size is a critical dimension in self-supervised learning. As shown in Ta-

Method	Pretrain data	Architecture	Frames	Top-1 (%)	Param(M)
OPN	UCF101	VGG	N/A	59.6	N/A
VCOP	UCF101	R(2+1)D	N/A	72.4	N/A
CoCLR	UCF101	S3D-G	32	81.4	9
VideoMAE	UCF101	ViT-B	16	90.8	87
MotionMAE-Sha	UCF101	ViT-B	16	94.0	87
SpeedNet	K400	S3D-G	64	81.1	9
Pace	K400	R(2+1)D	16	77.1	16
RSPNet	K400	S3D-G	64	93.7	9
ASCNet	K400	S3D-G	64	90.8	9
MMV	AS+HTM	S3D-G	32	92.5	9
XDC	IG65M	R(2+1)D	32	94.2	15
GDT	IG65M	R(2+1)D	32	95.2	15
VideoMAE	K400	ViT-B	16	96.1	87
MotionMAE-Sha	K400	ViT-B	16	96.3	87

Table 1: Comparison with the state-of-the-art methods on **UCF101**. Due to varying dataset sizes, we pretrain **MotionMAE** by 2400 epochs on UCF101 and by 1600 epochs on Kinetics-400 (K400). AS: Audio-Set [5]; HTM: HowTo100M [6]; IG65M: Instagram-65M [7]. N/A: Not Available. Grayed: Multimodal methods. Sha: Decoder with Sharing design.

ble 1, it is encouraging that our **MotionMAE** can surpass all the alternative methods in both domain-specific and domain-generic settings. For instance, when pretrained on Kinetics-400 (K400), our method reaches the best ever classification accuracy of **96.3%**, higher than the prior art self-supervised learning method VideoMAE [14] and multimodal learning method GDT [15] (despite using much more videos from IG65M, more video frames, and extra audio modality). This highlights the crucial significance of motion information which, once learned properly as in our proposed pretraining method, would demonstrate stronger representational power than other techniques (e.g., multimodal alignment). Another highlight is that in the domain-specific setting characterized by much lower training cost in this context, our **MotionMAE** achieving the top-1 accuracy of 94.0%, which is favored over the most similar competitor VideoMAE by as large as **3.2%**. These observations further verify the advantages of our method over prior alternatives.

Temporal Difference vs Optical Flow We investigate the effect of motion target by contrasting temporal difference (TD) with expensive-to-obtain optical flow (OF). We pre-extracted per-frame optical flow before pre-training. We use ViT-S as the backbone, pre-training for 400 epochs using **MotionMAE-Sha**, and finetuning for 50 epochs. Table 4 shows that both targets give very similar results. Thus TD is a more cost-effective choice.

Improvement vs Training Cost We also analyze the model accuracy and training time in GPU hours on Kinetics-400 and UCF101. For fair comparisons with prior art, we measure the latency under the same hardware setting consisting of 8 NVIDIA V100 GPUs. For Kinetics-400, we use pretraining (400 ~ 1600 epochs) and finetuning setting (50 epochs without repeat augmentation). For UCF101, we use pretraining (1600 ~ 3200 epochs) and finetuning setting (100 epochs without repeat augmentation). We test both variants (Sha and Ind) of **MotionMAE**. From Table 2(a) and Table 2(b), we observe that the findings of experiments are consistent with Something-something V2. In particular, **MotionMAE**

(a) Pretraining			
config	Something-Something V2	Kinetics400	UCF101
optimizer		AdamW [B]	
base learning rate		1.5e-4	
weight decay		0.05	
optimizer momentum		$\beta_1, \beta_2=0.9, 0.95$ [B]	
learning rate schedule		cosine decay [B]	
warmup epochs		40	
epochs	2400	1600	3200
flip augmentation	no	yes	yes
batch size	1024(B)512(L)	1024(B)512(L)	128(B)
augmentation		MultiScaleCrop	
patch norm	no	yes	yes
masking ratio	90%	90%	75%
masking type	random	tube	random
(b) Finetuning			
config	Something-Something V2	Kinetics400	UCF101
optimizer		AdamW [B]	
base learning rate	5e-4	5e-4(B)2e-3(L)	1e-3
weight decay		0.05	
optimizer momentum		$\beta_1, \beta_2=0.9, 0.999$	
layer-wise lr decay		0.75 [B]	
learning rate schedule		cosine decay [B]	
warmup epochs		5	
repeated sampling	2	2	1
epochs	30(B)20(L)	75(B)35(L)	100(B)
flip augmentation	no	yes	yes
batch size	512(B)256(L)	384(B)64(L)	256(B)
RandAug		(9, 0.5) [B]	
label smoothing		0.1 [B]	
mixup		0.8 [B]	
cutmix		1.0 [B]	
drop path		0.1(B),0.2(L)	

Table 2: Settings for model pretraining and finetuning. Note : $lr = base_lr \times batchsize / 256$ per the linear lr scaling rule.

(a) Kinetics-400.

Method	400			800			1600		
	Gain	Top-1	Time	Gain	Top-1	Time	Gain	Top-1	Time
VideoMAE	-	79.4	782.2	-	80.0	1564.4	-	80.5	3128.9
MotionMAE -Sha	+0.6	80.0	782.2	+0.4	80.4	1564.4	+0.2	80.7	3128.9
MotionMAE -Ind	+0.8	80.2	854.2	+0.3	80.3	1708.4	+0.4	80.9	3416.9

(b) UCF101.

Method	1600			2400			3200		
	Gain	Top-1	Time	Gain	Top-1	Time	Gain	Top-1	Time
VideoMAE	-	90.4	117.3	-	90.6	234.6	-	90.8	469.3
MotionMAE -Sha	+1.3	92.7	117.3	+3.0	93.6	234.6	+3.2	94.0	469.3
MotionMAE -Ind	+1.6	93.0	149.3	+2.8	93.4	298.6	+3.2	94.0	597.3

Table 3: Comparisons with the efficiency and effectiveness on Kinetics-400 and UCF101. We report the Gain, Top-1 Accuracy (%) and Training time. The Training time of pretraining is GPU hours.

Target	top-1 (%)	pre-process (days)
Frame	63.5	0
Frame + OF	64.3	2
Frame + TD	64.4	0

Table 4: Motion estimation: Temporal difference (TD) vs optical flow (OF) on Something-Something V2.

Masking strategy	top-1 (%)
time-only	76.2
random	79.8
space-only	80.0

Table 5: Effect of different masking strategies on Kinetics-400.

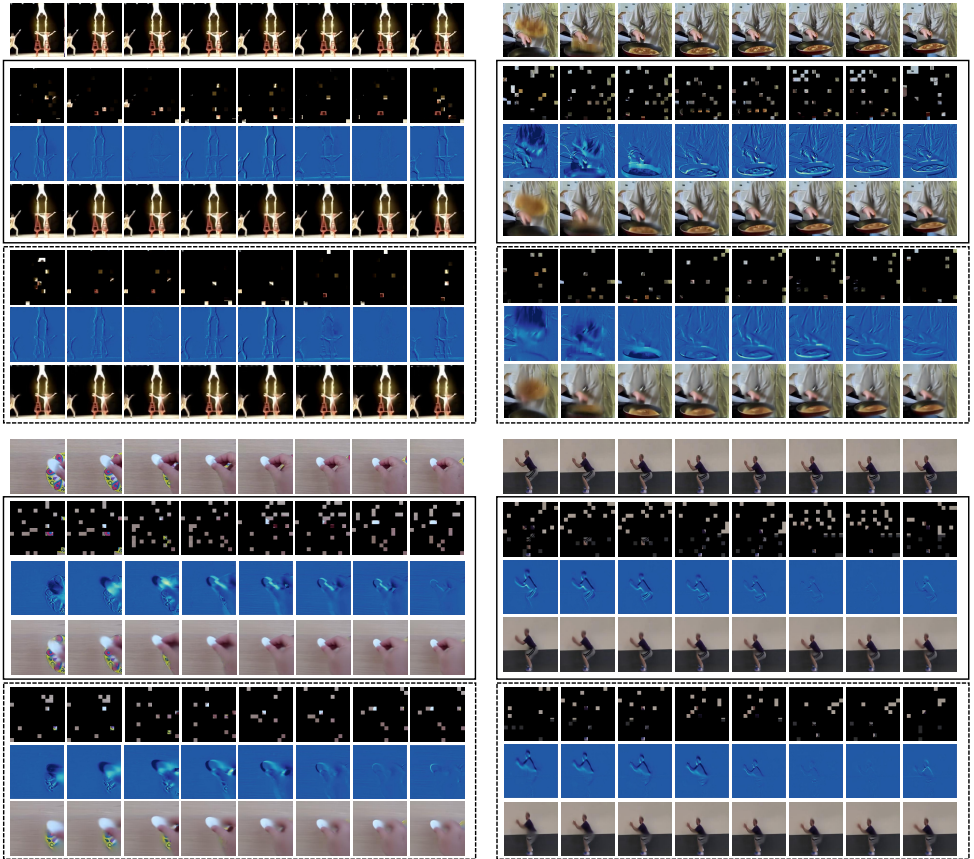


Figure 1: More visualizations on Kinetics-400 and UCF101. For each dataset, the *first* row shows the original video clip, and two *masking ratios* are visualized: **90%** (solid box) and **95%** (dashed box). Best viewed with zooming-in.

boosts more in the early epochs.

Mask sampling strategy We evaluate three masking strategies (random, time-only, space-only) on Kinetics400. As shown in Table 5, we find that random and space-only are similarly performing, whilst time-only is the worst. This is not surprising since reconstructing the whole frames could be over challenging.

C Additional Visualization Examples

Figure 1 and 2 provide more examples of reconstruction on Something-something V2, UCF101 and Kinetics-400.

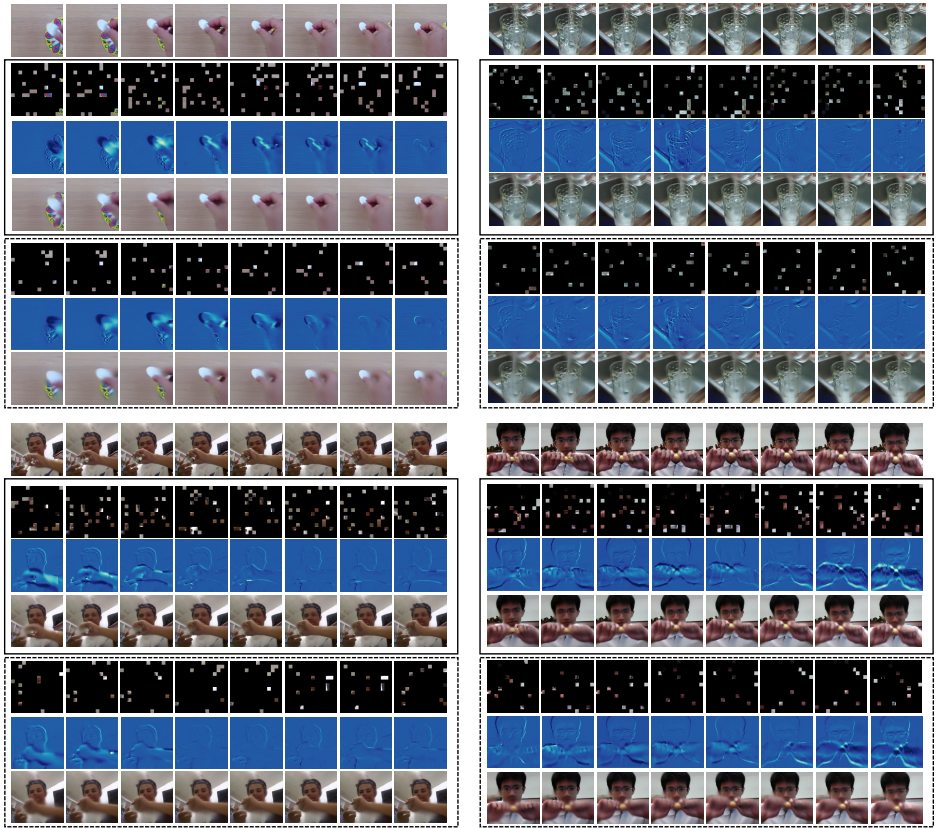


Figure 2: More visualizations on Something-something V2. The *first* row shows the original video clip, and two *masking ratios* are visualized: **90%** (solid box) and **95%** (dashed box). Best viewed with zooming-in.

References

- [1] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020.
- [2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations*, 2021.
- [3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.
- [4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

-
- [5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing*, pages 776–780. IEEE, 2017.
 - [6] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.
 - [7] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2016.
 - [8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International conference on machine learning*, 2017.
 - [9] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.
 - [10] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.
 - [11] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 2022.
 - [12] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.
 - [13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.