SELF-SUPERVISED VIDEO REPRESENTATION LEARNING WITH MOTION-AWARE MASKED AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked autoencoders (MAEs) have emerged recently as art self-supervised spatiotemporal representation learners. Inheriting from the image counterparts, however, existing video MAEs still focus largely on static appearance learning whilst are limited in learning dynamic temporal information hence less effective for video downstream tasks. To resolve this drawback, in this work we present a motion-aware variant – `MotionMAE`. Apart from learning to reconstruct individual masked patches of video frames, our model is designed to additionally predict the corresponding motion structure information over time. This motion information is available at the temporal difference of nearby frames. As a result, our model can extract effectively both static appearance and dynamic motion spontaneously, leading to superior spatiotemporal representation learning capability. Extensive experiments show that our `MotionMAE` outperforms significantly both supervised learning baseline and state-of-the-art MAE alternatives, under both domain-specific and domain-generic *pretraining*-then-*finetuning* settings. In particular, when using ViT-B as the backbone our `MotionMAE` surpasses the prior art model by a margin of 1.2% on Something-Something V2 and 3.2% on UCF101 in domain-specific pretraining setting. Encouragingly, it also surpasses the competing MAEs by a large margin of over 3% on the challenging video object segmentation task.

1 INTRODUCTION

Masked (denoising) autoencoding (MAE) (Vincent et al., 2010; 2008) has resurged as the state-of-the-art self-supervised image representation learning approach (He et al., 2022; Atito et al., 2021; Bao et al., 2021; Wei et al., 2022; Dosovitskiy et al., 2020; Li et al., 2021). This is inspired by the remarkable success of BERT (Devlin et al., 2018) in natural language processing. The idea of MAE is masked token/patch recovery within the Transformer framework (Vaswani et al., 2017). Existing image MAE models focus on visual tokenization Bao et al. (2021); Dong et al. (2021), token masking strategy (Li et al., 2021; Atito et al., 2021), reconstruction target (Wei et al., 2022), and architectural efficiency (He et al., 2022). Recently, MAE has been also applied to learn more challenging spatiotemporal representations from videos (Feichtenhofer et al., 2022; Tong et al., 2022; Wang et al., 2022). Despite demonstrating strong downstream task performances, we conjugate that these methods are still suboptimal in learning temporal structure information, since their learning objective is largely limited to the reconstruction of *individual tokens of video frames*, resulting in an over-challenging task of learning dynamic temporal information across frames.

To overcome the aforementioned limitation, in this work a simple yet superior video self-supervised learning method, dubbed as ***MotionMAE***, is introduced. Our idea is *local motion reconstruction* – recovering the motion information corresponding to each masked frame patch. The target motion information is obtained by simply contrasting two temporally adjacent video frames (*i.e.*, the temporal difference of video frames). This is an intrinsic temporal ingredient of video data, in addition to the spatial appearance of frame images. As illustrated in Fig. 1, the frame difference presents detailed motion information of push-up over time. Specifically, given a training video clip, `MotionMAE` randomly masks out a certain percentage (*e.g.*, 90%) of patches per frame, and reconstructs both masked frame patches for spacial appearance modeling and the corresponding motion structure for temporal dynamics modeling concurrently. For computational efficiency, we adopt an asymmetric

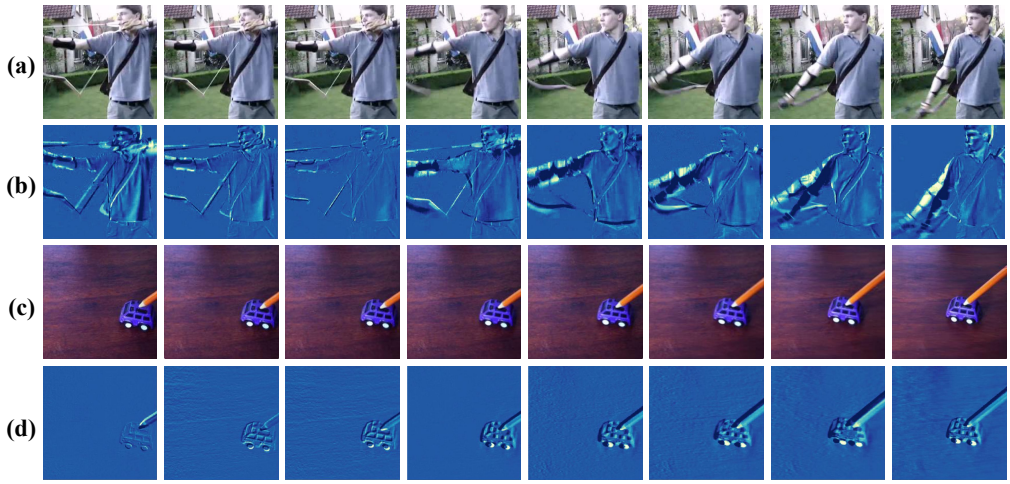


Figure 1: **Illustration of local motion information in video data**. Whilst **(a, c)** every single video frame provides rich appearance information, **(b, d)** their temporal difference gives additional dynamic motion knowledge over time. Intuitively, appearance and motion both are critical in learning spatiotemporal representation for video understanding.

MAE architecture (He et al., 2022) based on vanilla Vision Transformers (ViTs) (Dosovitskiy et al., 2020): The encoder operates only on visible patches and the decoder on all the patches.

In a nutshell, we propose a motion-aware MAE method, `MotionMAE`, for self-supervised spatiotemporal representation learning from unlabeled videos. For extensive evaluation, we conduct both domain-generic and domain-specific *pretraining*-then-*finetuning* paradigms. The results show that our `MotionMAE` outperforms both supervised learning and alternative state-of-the-art MAE designs on action classification, often by a large margin. In particular, `MotionMAE` achieves an accuracy of 75.5% on Something-Something V2, 85.3% on Kinetics-400, and 94.0% on UCF101 under domain-specific pretraining setting. The performance advantage of our model remains on the challenging Video Object Segmentation task with a margin of 3+% over the competing MAE models on the DAVIS2017 benchmark.

2 Related Work

Self-supervised learning of video representation Self-supervised learning (SSL) of visual representation (He et al., 2020; Grill et al., 2020; Chen et al., 2020b; Caron et al., 2020) have advanced massive recently and demonstrated superior performance over previously dominant supervised learning (*e.g.*, on ImageNet with exhaustive manual labeling). This opens up a door to achieve ever stronger representations due to the availability of much larger volume of (even infinite) unlabeled visual data from diverse sources. This inspires SSL for video data with extra temporal dimension as compared to static images. In the literature, the previous waive of video SSL methods rely on carefully designing time-related pretext tasks in three lines: (1) Predicting specific transformations (*e.g.*, rotation angle (Jing et al., 2018), playback speed (Benaim et al., 2020), temporal order (Misra et al., 2016; Lee et al., 2017; Xu et al., 2019) and motion statistics (Wang et al., 2019)); (2) Predicting future frame (Han et al., 2019; 2020a); (3) Instance discrimination (Qian et al., 2021; Wang et al., 2020; Chen et al., 2021). Usually, top-performing methods need to conduct strong and diverse data augmentations, which brings about the notorious scalability bottleneck and learning complexity. Also, the model performance is conditioned on a good number of tricks (*e.g.*, feature bank, data augmentation design).

Masked autoencoders With the increasingly wide adoption of Transformers (Vaswani et al., 2017) in computer vision, masked autoencoders (MAEs) have recently emerged as a general SSL framework (He et al., 2022; Bao et al., 2021; Radford et al., 2018). This is mostly inspired by the success of BERT (Devlin et al., 2018) in a wide range of NLP downstream tasks. In general, it can

be also regarded as a special pretext task. However, compared to most previous alternatives, MAEs are not only simpler in design, but also stronger in representation learning.

Specifically, image MAE variants learn the representation by predicting the masked/unknown regions from visible parts. For example, iGPT (Chen et al., 2020a) operates on sequences of pixels and predicts unknown pixels. ViT (Dosovitskiy et al., 2020) predicts the mean colors of masked patches. BEiT (Bao et al., 2021) further improves MAEs performance with masked visual token prediction. PeCo (Dong et al., 2021) suggests to inject perceptual similarity during visual codebook learning. Interestingly, MAE (He et al., 2022) demonstrates the strength of the straightforward idea of image patch reconstruction, in addition to improving the pretraining efficiency by adopting high masking ratios and encoding only unmasked patches. Alternatively, MaskFeat (Wei et al., 2022) leverages HOG (Dalal & Triggs, 2005) feature as the prediction target to yield strong visual representation.

Very recently, MAE has been similarly applied for learning more challenging spatiotemporal representations from videos. For example, by extending the BEiT pretraining paradigm, BEVT (Wang et al., 2022) decouples masked SSL into spatial representation learning on images and temporal dynamics learning on both images and videos in a two-stage process. VideoMAE (Tong et al., 2022) and MAE (Feichtenhofer et al., 2022) simply reconstruct masked spatiotemporal patches of each video against even high masking ratios (*e.g.*, 90%) and achieve strong performance on video downstream tasks. This result is encouraging and inspiring given their simplicity. However, we consider that such straight inheriting from image MAEs would be limited in learning temporal structure information. This is because their learning objective tends to focus on reconstruction of *individual tokens of video frames*, since leaving the model itself to derive the underlying implicit spatiotemporal representation is drastically challenging. To tackle this problem, we present a motion-aware MAE that additionally predicts the corresponding motion structure information. The target motion to be reconstructed is obtained by contrasting two temporally adjacent video frames (*i.e.*, the temporal difference of video frames). Extensive experiments and analysis validate our proposed method.

3 METHOD

Our proposed `MotionMAE` adopts the asymmetric Transformer based masked autoencoder architecture (He et al., 2022), as depicted in Fig. 2. The objective is to *pretrain a video representation model* in a self-supervised manner for facilitating downstream tasks such as action classification. Next, we describe the details of our model architecture and pretraining.

Patch embedding We divide evenly an input video into a regular grid of non-overlapping space-time patches (*i.e.*, cubes) and perform cube embedding (Bertasius et al., 2021; Arnab et al., 2021). This reduces the temporal resolution of video to suppress the redundancy degree over time. Learnable positional embeddings are applied via elementwise addition.

Patch masking We randomly mask the patches at each time point of a video clip at a fixed ratio. Common masking strategies include random, space-only (*i.e.*, tube), and time-only (masking the whole frames selected). As video data present more redundancy than both text and image, setting a higher masking ratio (*e.g.*, 90%) is helpful.

Autoencoding We adopt a vanilla ViT (Dosovitskiy et al., 2020) as the encoder with space-time joint attention. Following (He et al., 2022; Feichtenhofer et al., 2022), the encoder is applied only on unmasked patches. This greatly reduces the time and memory complexity, leading to a more efficient solution. The encoder can capture high-level spatiotemporal semantic information by interacting patches of all frames.

Our decoder is another vanilla ViT deployed on the union of the encoded patch set and a set of mask tokens (He et al., 2022). It consists of two heads in parallel: ***a space head*** to predict masked frame patches, and ***a time head*** to predict the corresponding local motion structure. The motion target is obtained by computing L_1 difference of pixel values between two temporally nearby frames, which encodes the short-term temporal dynamics information underlying across frames. This dynamics information underlies along the time dimension which is complementary to the appearance information of individual video frames. Our motivation is that by decomposing the two types of information

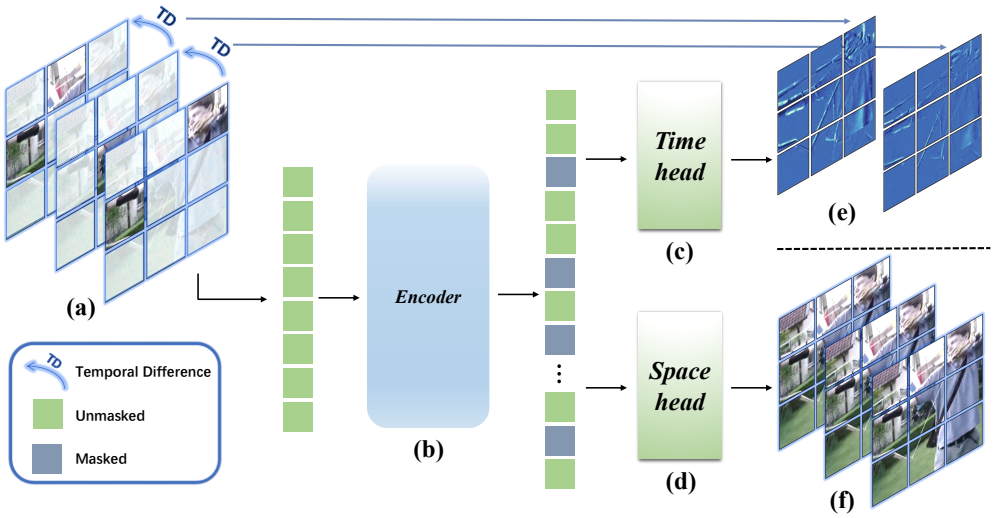


Figure 2: Illustration of the proposed `MotionMAE`. **(a)** We randomly mask the patches of each frame of a given video clip at a fixed ration. In *pretraining*, **(b)** an encoder operates on the set of visible patches. **(c)** A time head and **(d)** a space head then processes the full set of encoded patches and mask tokens to reconstruct **(e)** the masked frame patches and **(f)** the corresponding local motion, concurrently. After pretraining, we discard the time head and space head then evaluate the pretraining quality of the encoder on downstream task.

underlying the video data, self-supervised learning of spatiotemporal representation could be facilitated. This could be considered as introduction of prior knowledge which otherwise needs to be discovered from the learning process itself. For each decoder head, specific positional embeddings are applied following MAE (He et al., 2022). Since video data is high-dimensional, the decoder is designed to be more lightweight than the encoder, consuming less computational cost despite processing the full patch set with two heads.

Objective loss function For model pretraining, for each head we minimize the L_2 reconstruction loss between the prediction and the target (*i.e.*, the frame image patches for the space head and frame difference patches for the time head) over the masked patches. The final objective is to minimize their summation.

4 Implementation

Downstream tasks We focus on two significant downstream tasks: action recognition, and video object segmentation. For action recognition, we evaluate three datasets: UCF101 (Soomro et al., 2012), Kinetics-400 (Kay et al., 2017), and Something-Something V2 (Goyal et al., 2017). Specifically, *Kinetics-400* contains around 240k training videos and 20k validation videos at the length of 10 seconds from 400 action classes. *Something-Something V2* offers around 169k videos for training and 20k videos for validation. In contrast to Kinetics-400, this dataset contains 174 motion-centric action classes. These two large-scale video datasets focus on different visual cues for action recognition. *UCF101* is a small video dataset with around 9.5k/3.5k train/val videos. This allows for testing with limited training data. All the three datasets together provide a cohort with varying-sized benchmarks for extensive evaluation.

For video object segmentation, we test the *DAVIS-2017* (Pont-Tuset et al., 2017) dataset with 30 videos for validation and 59 object classes. The task is to segment the salient moving objects in the scene.

Data pre-processing For MotionMAE pretraining, by default the input is a clip of 16 frames at a resolution of 224×224 pixels (except otherwise stated). A training clip is sampled from a raw video. We set the temporal stride of 2, 4, and 4 for Something-Something V2, Kinetics-400, and UCF101, respectively, with the starting frame randomly chosen. For the spatial dimensions, we perform random resized cropping at a scale range of [0:5; 1]. During both pretraining and finetuning, we apply flip augmentation on Kinetics-400 and UCF101, but not on Something-Something V2 as this dataset involves order-sensitive action classes. We do not apply any other data augmentations.

Architecture We adopt the vanilla ViT architecture (Dosovitskiy et al., 2020) as the encoder and decoder, namely the ViT-B and ViT-L variants. For a patch, we use a temporal size of 2 and a spatial size of 16 ×16, denoted as 2 ×16 ×16. Given a 16 ×224 ×224 input, the corresponding patch size is 8 ×14 ×14 tokens. We do not use the [CLS] token in the ViT models, yielding a small improvement in running-time without performance loss. We use a Transformer decoder consisting of 4 layers with 384 and 512 embedding dimensions for ViT-B and ViT-L, respectively. The decoder outputs both frame patches and motion patches at two heads. We use sinusoidal positional encoding in both the encoder and the decoder.

Masking During pretraining, similar as prior works (Tong et al., 2022; Feichtenhofer et al., 2022), we apply a masking ratio of 90% for large datasets (*e.g.*, Kinetics-400 and Something-Something V2) and 75% for small ones (*e.g.*, UCF101). We find the masking ratio is partly conditioned on the dataset scale. For the masking strategy, we find the best option for UCF101, Something-Something V2 and Kinetics400 to be random, random and space-only (*i.e.*, tube), respectively. In general, space-only and random can yield similar performance, whist time-only performs the worst. We ablate the masking hyper-parameters and strategies in the appendix.

Setting For fair comparison, we follow the pretraining configuration of (Tong et al., 2022). Specifically, we use the AdamW optimizer with a batch size of 1024. We evaluate the pretraining quality by *end-to-end fine-tuning*. In model inference, we adopt the common multi-view practice: (1) Taking K temporal clips ($K = 2, 5, 10$ for Something-Something V2, Kinetics-400 and UCF101) to cover the whole video length, and (2) for each clip, taking 3 spatial views to cover the longer spatial axis, resulting in a total of $3K$ views. The final prediction is obtained by averaging all the views. More implementation details and hyper-parameters are given in the appendix.

5 Experiments

5.1 Action Recognition

We contrast the proposed MotionMAE with two categories of state-of-the-art video representation learning methods: (1) **Supervised video learning methods** including SlowFast (Feichtenhofer et al., 2019), TDN (Wang et al., 2021), Timesformer (Bertasius et al., 2021), ViViT (Arnab et al., 2021), MViT (Fan et al., 2021; Li et al., 2022), Video Swin (Liu et al., 2022) and MotionFormer (Patrick et al., 2021). (2) **Self-supervised video learning methods** including OPN (Lee et al., 2017), VCOP (Xu et al., 2019), CoCLR (Han et al., 2020b), SpeedNet (Benaim et al., 2020), Pace (Wang et al., 2020), RSPNet (Chen et al., 2021), ASCNet (Huang et al., 2021), MMV (Alayrac et al., 2020), XDC (Alwassel et al., 2020), GDT (Patrick et al., 2020), BEVT (Wang et al., 2022), Maskfeat (Wei et al., 2022), VideoMAE (Tong et al., 2022) and MAE (Feichtenhofer et al., 2022).

For more extensive comparison, we consider both domain-generic (*i.e.*, training and test on different datasets) and domain-specific (*i.e.*, training and test on the same dataset) pretraining-then-finetuning settings.

Something-Something-V2 On this temporal structure sensitive dataset, from Table 1 we draw several observations. **(1)** In terms of learning strategy, self-supervised learning is generally superior over the conventional supervised-learning. For example, MViTV2 (Li et al., 2022) is clearly outnumbered by MaskFeat (Wei et al., 2022) despite using extra data (IN21K). **(2)** Domain-specific pretraining is often more performing than the domain-generic counterpart as reflected in the contrast of VideoMAE (Tong et al., 2022) and MAE (Feichtenhofer et al., 2022). **(3)** In either setting, our MotionMAE exceeds all the alternative solutions by a descent margin. Specifically, our method out-

Method	Extra data	Backbone	Frames	Top-1 (%)	GFLOPs	Param(M)
SlowFast	K400	SlowFast	8+32	63.1	106×1×3	63.1
TDN	IN1K	ResNet101	8+16	69.6	198×3×1	88
TimeSformer	K400+IN21K	ViT-L	64	62.4	5549×1×3	430
ViViT	K400+IN21K	ViT-L	32	65.9	995×4×3	N/A
MViTv1	K400	MViT-B	64	67.7	455×3×1	37
MotionFormer	K400+IN21K	ViT-L	32	68.1	1185×1×3	382
Video Swin	K400+IN21K	Swin-B	32	69.6	321×1×3	88
MViTv2	K400 + IN21K	MViTv2-L	40	73.3	2828×1×3	213
BEVT	K400 + IN1K	Swin-B	32	71.4	321×1×3	88
MaskFeat	K400	MViTv2-L	40	74.4	2828×1×3	218
MAE	K400	ViT-L	16	72.1	598×1×3	304
MAE	K400	ViT-H	16	74.1	1193×1×3	632
VideoMAE	-	ViT-B	16	70.6	180×2×3	87
VideoMAE	-	ViT-L	16	74.2	598×2×3	305
VideoMAE	-	ViT-L	32	75.3	1436×2×3	305
MotionMAE	K400	ViT-L	16	**74.3**	598×2×3	305
MotionMAE	-	ViT-B	16	**71.8**	180×2×3	87
MotionMAE	-	ViT-L	16	**74.6**	598×2×3	305
MotionMAE	-	ViT-L	32	**75.5**	598×2×3	305

Table 1: Comparison with the state-of-the-art methods on **Something-Something V2**. Due to varying dataset sizes, we pretrain our `MotionMAE` by 1600 epochs on Kinetics-400 (K400) and by 2400 epochs on Something-Something V2. `IN21K`: ImageNet-21K; `IN1K`: ImageNet-1K. GFLOPs format: A single view's × temporal views × spatial views. `N/A`: Not Available. Grayed: Supervised learning methods.

Method	Extra data	Backbone	Frames	Top-1 (%)	GFLOPs	Param(M)
SlowFast	IN1K	SlowFast	16+64	79.8	234×10×3	60
TDN	IN1K	ResNet101	8+16	79.4	198×10×3	88
TimeSformer	IN21K	ViT-L	96	80.7	8353×1×3	430
ViViT	IN21K	ViT-L	128	81.7	3980×3×1	N/A
MViT	-	MViT-B	64	81.2	455×3×3	37
MotionFormer	IN21K	ViT-L	32	80.2	1185×10×3	382
Video Swin	IN21K	Swin-L	32	83.1	604×1×3	197
MViTv2	IN21K	MViTv2-B	32	82.9	255×1×3	213
BEVT	IN1K	Swin-B	32	81.1	282×4×3	88
MaskFeat	-	MViTv2-L	40	84.3	377×10×1	218
VideoMAE	-	ViT-B	16	81.5	180×5×3	87
VideoMAE	-	ViT-L	16	85.1	598×5×3	305
MAE	-	ViT-L	16	84.8	598×7×3	304
MAE	-	ViT-H	16	85.1	1193×7×3	632
MotionMAE	-	ViT-B	16	**81.7**	180×5×3	87
MotionMAE	-	ViT-L	16	**85.3**	598×5×3	305

Table 2: Comparison with the state-of-the-art methods on **Kinetics-400** (K400). We pretrain our `MotionMAE` by 1600 epochs on Kinetics-400 (K400). `IN21K`: ImageNet-21K; `IN1K`: ImageNet-1K. GFLOPs format: A single view's × temporal views × spatial views. `N/A`: Not Available. Grayed: Supervised learning methods.

performs VideoMAE by a range of $[0.2\%, 1.2\%]$ in the domain-specific setting and MAE by 2.2% in domain-generic setting. Besides, ViT-L based `MotionMAE` achieves the unprecedented top-1 accuracy of **75.5 %**. This validates the advantage of our pretraining method in learning spatiotemporal representation from unlabeled video data.

Kinetics-400 Unlike Something-Something V2, appearance information plays a more dominant role in Kinetics-400. However, from Table 2 we can still see similar observations. For example, **(1)** `MotionMAE` remains the best pretraining method among all the competitors. This suggests that

Method	Extra data	Architecture	Frames	Top-1 (%)	Modality	Param(M)
OPN	-	VGG	N/A	59.6	V	N/A
VCOP	-	R(2+1)D	N/A	72.4	V	N/A
CoCLR	-	S3D-G	32	81.4	V	9
SpeedNet	K400	S3D-G	64	81.1	V	9
Pace	K400	R(2+1)D	16	77.1	V	16
RSPNet	K400	S3D-G	64	93.7	V	9M
ASCNet	K400	S3D-G	64	90.8	V	9M
MMV	AS+HTM	S3D-G	32	92.5	V+A+T	9
XDC	IG65M	R(2+1)D	32	94.2	V+A	15
GDT	IG65M	R(2+1)D	32	95.2	V+A	15
VideoMAE	-	ViT-B	16	90.8	V	305
VideoMAE	K400	ViT-B	16	96.1	V	305
MotionMAE	-	ViT-B	16	**94.0**	V	87
MotionMAE	K400	ViT-B	16	**96.3**	V	305

Table 3: 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 (Gemmeke et al., 2017); `HTM`: HowTo100M (Miech et al., 2019); `IG65M`: Instagram-65M (Ghadiyaram et al., 2019). `V`: Visual; `A`: Audio; `T`: Text or narration; `N/A`: Not Available. Grayed: Multimodal methods.

Method	Pretrain data	Architecture	J&F-Mean	J-Mean	F-Mean
MAE	IN1K	ViT-B	50.9	49.3	52.6
MAE	K400	ViT-B	53.5	52.6	54.4
VideoMAE	K400	ViT-B	53.8	53.2	54.4
MotionMAE	K400	ViT-B	**56.8**	**55.8**	**57.8**

Table 4: Video object segmentation results on DAVIS 2017. The pretrained models are frozen *without* task-specific finetuning. `Metrics`: mean region similarity J-Mean and mean contour-based accuracy F-Mean. `IN1K`: ImageNet-1K; `K400`: Kinetics-400.

dedicated motion recovery from the masking as we introduce is also beneficial. **(2)** By achieving the top-1 accuracy of **85.3%**, ViT-L based `MotionMAE` even surpasses ViT-H based MAE (85.1%), while enjoying high efficiency in GFLOPs. These results indicate the generic performance superiority of our method over diverse action genres.

UCF101 Following the two 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 Table 3, 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-suerpvised learning method VideoMAE (Tong et al., 2022) and multimodal learning method GDT (Patrick et al., 2020) (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 congtext, our `MotionMAE` is favored over the most similar competitor VideoMAE by as large as **3.2%**. These observations further verify the rich advantages of our method over existing alternatives.

5.2 Video Object Segmentation

In addition to action recognition evaluation, next we evaluate the usefulness of our method in a task of segmenting salient objects in videos. This presents a different testing perspective as it is highly object-centered. To that end, we compare `MotionMAE` with two prior art self-supervised video learners (MAE (Feichtenhofer et al., 2022) and VideoMAE (Tong et al., 2022)). We freeze the pretrained models to purely evaluate their expressiveness of video objects over space and time, *without*

Target	top-1 (%)
Frame	67.8
Motion	67.6
Frame+Motion	**68.4**

(a) Reconstruction target.

Ratio	top-1 (%)
1:2	68.2
1:1	**68.4**
2:1	68.3

(b) Weight ratio: frame to motion.

MG	top-1 (%)
1	**68.4**
2	67.8
3	67.4

(c) Motion granularity (MG).

Decoder width	top-1 (%)
128	67.5
384	**68.4**
512	68.3

(d) Decoder width.

Decoder depth	top-1 (%)
1	67.4
4	**68.4**
8	68.2

(e) Decoder depth.

Loss function	top-1 (%)
L_1	67.6
MSE	**68.4**
Smooth L_1	67.5

(f) Reconstruction loss.

Table 5: Ablation experiments on **Something-Something V2** under the domain-specific pretraining setting. Pretraining length of our `MotionMAE`: 400 epochs. `Model`: ViT-B with the input video clip at a resolution of 16×224×224.

any finetuning. We adopt the same experimental protocol as in (Jabri et al., 2020) by segmenting the scenes with a nearest-neighbor between consecutive frames. We draw a couple of observations from Table 4: **(1)** Pretraining data makes difference as exampled by MAE, and video data is generally more effective as expected. This suggests the importance of learning spatiotemporal representation to this task. **(2)** When the same video dataset is used, our method surpasses both competing MAEs by a margin of 3.0% or more. This again validates the superiority of our video representation learning method in the object-focused scenario, thanks to the proposed masked motion information modeling in addition to static appearance information reconstruction.

5.3 Ablation Study

We perform ablation study on Something-Something V2 in the most-performing domain-specific pretraining setting. We use ViT-B as the backbone under the inference setting of 2 clips × 3 crops.

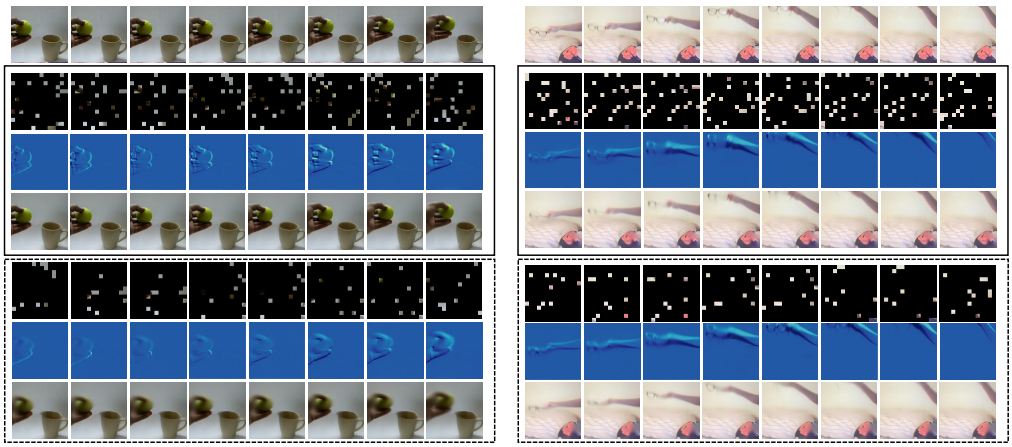
Reconstruction target In this study, we conjugate that a key ingredient with MAE is reconstruction target. To evaluate this, we swap the reconstruction target among the choices of {`frame`, `motion (temporal difference)`, `frame+motion`)}, whilst keeping all the others. We summarize a few key observations: **(1)** As shown in Table 5a, the `frame` and `motion` targets are similarly effective, which is somewhat out of expectation. **(2)** When they are combined, a clear performance gain can be obtained, suggesting their good complementing effect. Further, as shown in Table 5a, their combining weight is only slightly sensitive to the performance. **(3)** Regarding the motion temporal granularity, the results in Table 5c verify that the most fine-grained motion (*i.e.*, obtained by contrasting the directly adjacent frames) is most useful, as expected. A plausible reason is that predicting large-time-gap motion is over challenging due to more complex dynamic variations over time.

Decoder design Another key component in MAE is the design of decoder. To test this, we vary the capacity of our decoder in terms of width and depth. As shown in Table 5d and Table 5e, it is important that both the width (*e.g.*, 384 dimensions) and depth (*e.g.*, 4 blocks) need to be sufficiently large but not too many.

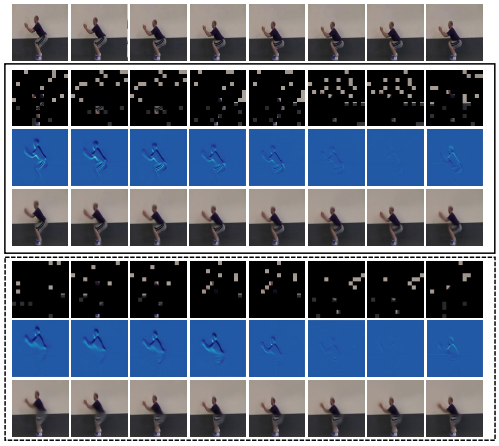
Reconstruction loss function Lastly, we evaluate the effect of reconstruction loss function. We consider three options: L_1, `Smooth` L_1, and Mean Squared Error (`MSE`). As shown in Table 5f, `MSE` yields the best result among these.

5.4 Model Behavior Visualization

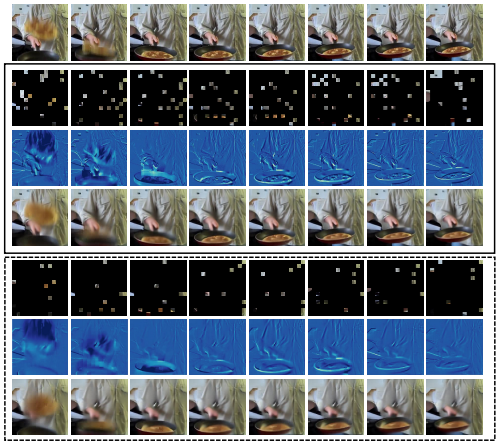
To examine the model pretraining behavior, we visualize the reconstruction output. As shown in Figure 3, under 90% masking, our `MotionMAE` model can well generalize the reconstruction ability to new video samples for both individual video frames and dynamic inter-frame motion. When



(a) Something-something V2



(b) UCF101



(c) Kinetics-400

Figure 3: Example reconstruction of our `MotionMAE` on the validation set of (a) Something-Something v2, (b) UCF101, and (c) Kinetics-400. 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.

increasing the masking ratio to 95%, the reconstruction results still preserve the most information about both appearance and motion although less fine-grained. This verifies that our model can recover well the masked data to support proper self-supervised learning of video representation.

6 Conclusion

In this paper, we have presented motion-aware Masked Autoencoders, namely `MotionMAE`, for self-supervised video representation learning. Apart from learning to reconstruct individual masked patches of video clips, our model is designed particularly to concurrently predict the corresponding motion structure information over time. Compared to prior art MAEs, our model can better perceive both static appearance and dynamic motion spontaneously, yielding superior spatiotemporal representation learning. Empirical results on standard action recognition benchmarks demonstrate that our `MotionMAE` outperforms significantly both supervised learning baseline and state-of-the-art MAE alternatives, under both domain-specific and domain-generic *pretraining*-then-*finetuning* settings. The generalization and transferability of our pretrained models have been also evaluated on another challenging motion-centric task, video object segmentation.

REFERENCES

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

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.

Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9922–9931, 2020.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International conference on machine learning*, volume 2, pp. 4, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

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*, pp. 1691–1703. PMLR, 2020a.

Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1045–1053, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

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*, pp. 702–703, 2020.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition*, volume 1, pp. 886–893. Ieee, 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

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*, pp. 776–780. IEEE, 2017.

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*, pp. 12046–12055, 2019.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 312–329. Springer, 2020a.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020b.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8096–8105, 2021.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pp. 667–676, 2017.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022.

Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International conference on machine learning*, 2017.

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*, pp. 2630–2640, 2019.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544. Springer, 2016.

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.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6964–6974, 2021.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4006–4015, 2019.

Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pp. 504–521. Springer, 2020.

Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1895–1904, 2021.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14733–14743, 2022.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14668–14678, 2022.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10334–10343, 2019.

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*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

A Additional Implementation Details

config	Something-Something V2	Kinetics400	UCF101
optimizer	AdamW (Loshchilov & Hutter, 2017)		
base learning rate	1.5e-4		
weight decay	0.05		
optimizer momentum	β_1, β_2=0.9, 0.95 (Chen et al., 2020a)		
learning rate schedule	cosine decay (Loshchilov & Hutter, 2016)		
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

(a) Pretraining

config	Something-Something V2	Kinetics400	UCF101
optimizer	AdamW (Loshchilov & Hutter, 2017)		
base learning rate	5e-4	5e-4(B)2e-3(L)	1e-3
weight decay	0.05		
optimizer momentum	β_1, β_2=0.9, 0.999		
layer-wise lr decay	0.75 (Bao et al., 2021)		
learning rate schedule	cosine decay (Loshchilov & Hutter, 2016)		
warmup epochs	5		
repeated sampling	2	2	1
epochs	30(B)20(B)	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) (Cubuk et al., 2020)		
label smoothing	0.1 (Asano et al., 2020)		
mixup	0.8 (Zhang et al., 2017)		
cutmix	1.0 (Yun et al., 2019)		
drop path	0.1(B),0.2(L)		

(b) Finetuning

Table 6: Settings for model pretraining and finetuning. `Note`: *lr* = *base_lr*×batchsize / 256 per the linear *lr* scaling rule.

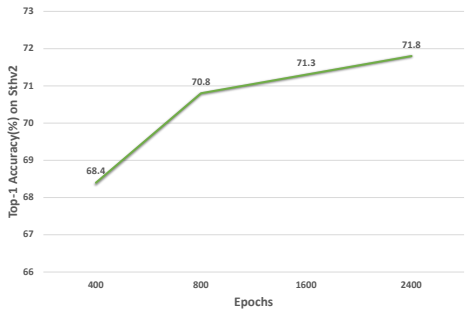
We pretrain `MotionMAE` on Something-Something V2, Kinetics-400 and UCF101 using the hyper-parameters as summarized in Table 6a. The dataset specific hyper-parameters are given in the individual columns, with the others shared across datasets. These settings apply to ViT-B and ViT-L, unless specified otherwise. We use the same hyper-parameters for pretraining. Table 6b 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.

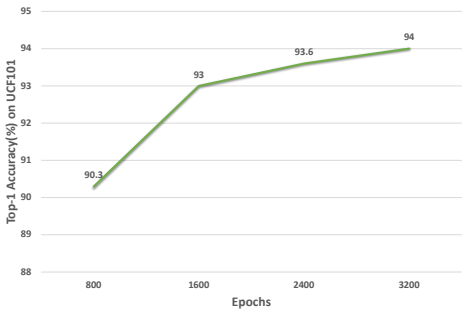
B Additional Experimental Results

Training schedule In Figure 4 we show the effect of training schedule on Something-Something (large) and UCF101 (small). It is clear that longer training schedule brings slightly gains on both datasets. More training might lead to further improvement at extra computation cost. This indicates high potential of our model.

Local motion extraction: temporal difference vs. optical flow Except temporal difference (TD) as we have exploited, optical flow is another popular method for local motion estimation. It is



(a) Performance on Something-something V2



(b) Performance on UCF101

Figure 4: The effect of training schedules on (a) Something-Something V2 and (b) UCF101.

Target	top-1 (%)
Frame	93.1
Frame + OF	93.5
Frame + TD	**94.0**

Table 7: Motion estimation: temporal difference (TD) and optical flow (OF) on UCF101.

Masking strategy	top-1 (%)
time-only	74.6
random	78.8
space-only	**79.0**

Table 8: Effect of different masking strategies on Kinetics-400.

Ratio	UCF	Something-Something V2
75%	**94.3**	66.0
90%	92.7	**68.4**

Table 9: Effect of masking ratio.

Arch	top-1 (%)
Att. share	68.0
Divide	**68.4**

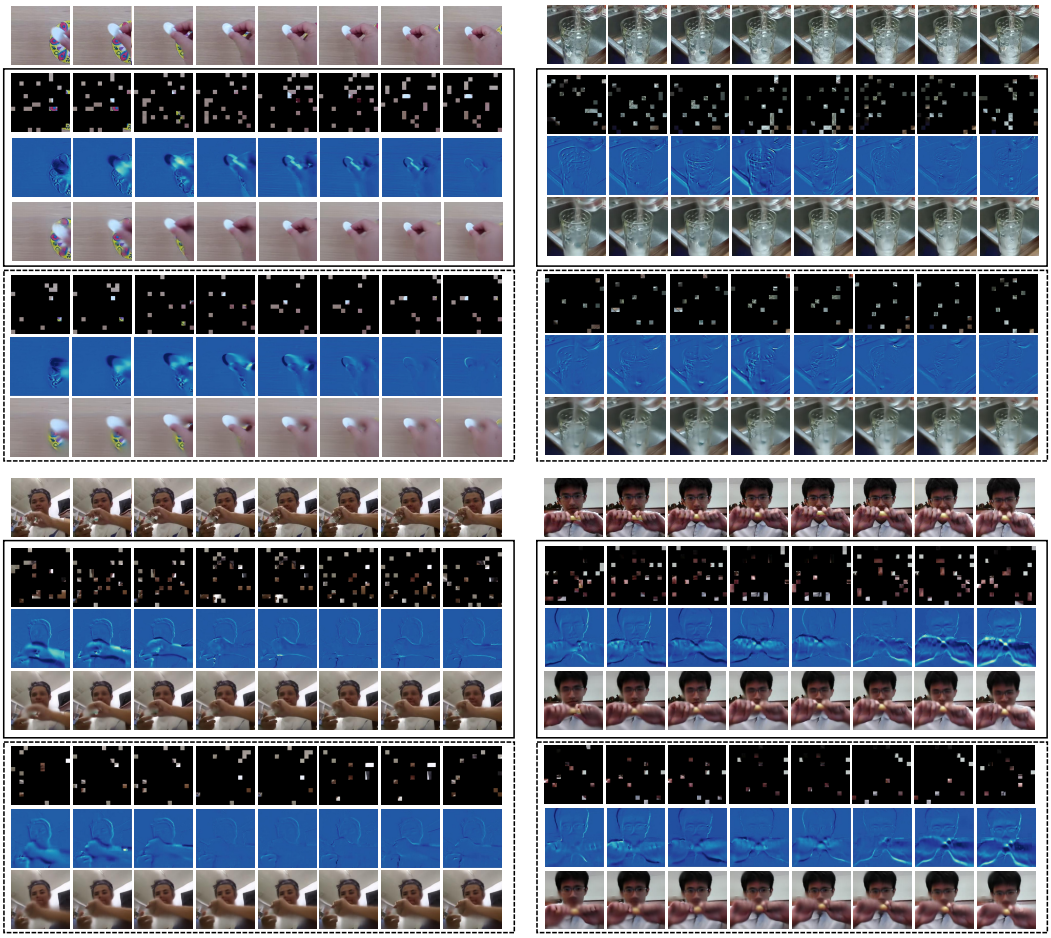
Table 10: Effect of decoder architecture on Something-Something V2.

interesting to see which one is better and why. We conduct this test on UCF101. Specifically, we pre-extract per-frame optical flow (OF) to replace the TD during model pre-training, whilst keeping the others the same. As shown in Table 7, we can see that OF is relatively inferior to TD, whilst still improve over using frames only. This is seemingly out of expectation since OF is often considered to be more accurate motion estimation at much more computational cost. We consider that the underlying reason is that the frames used to compute OF is not fully available in the input. Suppose the t-th frame is sampled to a video clip, the $t+1$-th frame which is used to compute OF, is however excluded by our sampling strategy (see `data pre-processing` in Section 4). This discrepancy may cause the lower performance by OF. In contrast, our TD has no such problem whilst being more efficient.

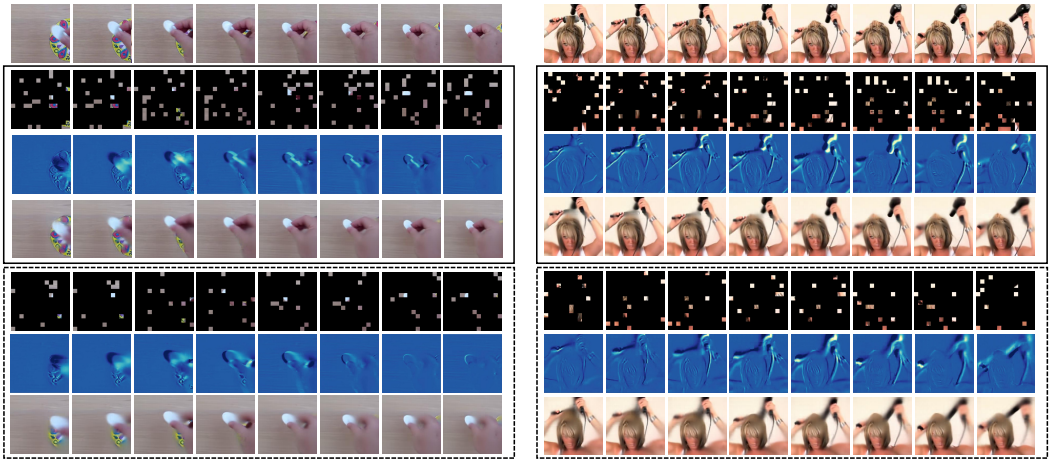
Mask sampling strategy We evaluate three masking strategies (random, time-only, space-only) on Kinetics400. As shown in Table 8, we find that random and space-only are similarly performing, whilst time-only is the worst. This is not surprising since reconstructing the whole frames could be over challenging.

Masking ratio We evaluate the effect of mask ratio on Something-something V2 (large) and UCF101 (small). We adopt the data-specific pretraining setting. As shown in Table 9, this setting is considerably dataset (*e.g.*, size) specific. This is not observed in the previous works.

Decoder architecture We evaluate the effect of decoder architecture design. By default, we adopt a *parallel* structure with two independent networks: a space head to predict masked frame patches, and a time head to predict the corresponding local motion structure. An alternative is to *share* a network except the last prediction layer. As shown in Table 10, network sharing is less effective, confirming our design choice.



(a) Something-something V2



(b) UCF101

Figure 5: More visualizations on Something-something V2 and UCF101 following Figure 3.

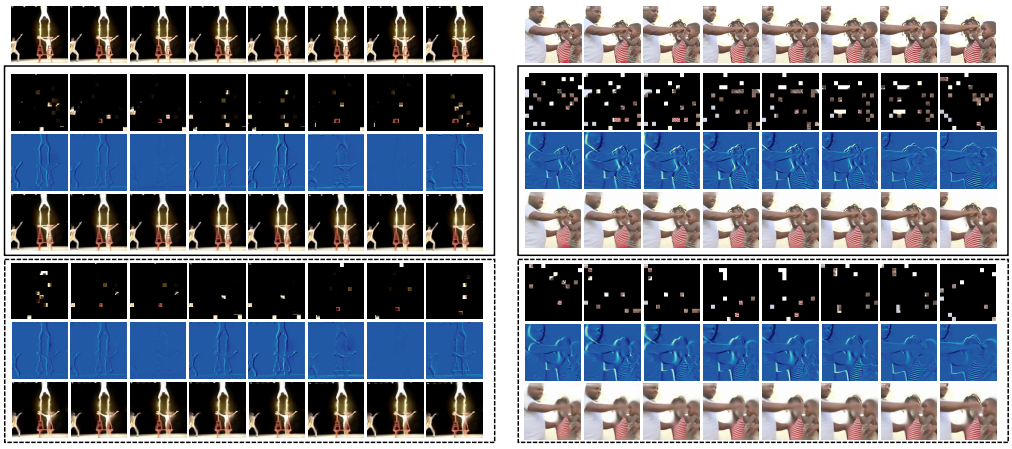


Figure 6: More visualizations on Kinetics-400 following Figure 3.

C ADDITIONAL VISUALIZATION EXAMPLES

Figure 5 and 6 illustrates more examples of reconstruction on Something-something V2, UCF101 and Kinetics-400.