Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning

Rui Wang[1,2]   Dongdong Chen[3]   Zuxuan Wu[1,2†]   Yinpeng Chen[3]   Xiyang Dai[3]

Mengchen Liu[3]   Lu Yuan[3]   Yu-Gang Jiang[1,2†]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University

[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing

[3]Microsoft Cloud + AI

Abstract

*Benefiting from masked visual modeling, self-supervised video representation learning has achieved remarkable progress. However, existing methods focus on learning representations from scratch through reconstructing low-level features like raw pixel values. In this paper, we propose masked video distillation (MVD), a simple yet effective two-stage masked feature modeling framework for video representation learning: firstly we pretrain an image (or video) model by recovering low-level features of masked patches, then we use the resulting features as targets for masked feature modeling. For the choice of teacher models, we observe that students taught by **video teachers** perform better on temporally-heavy video tasks, while **image teachers** transfer stronger spatial representations for spatially-heavy video tasks. Visualization analysis also indicates different teachers produce different learned patterns for students. To leverage the advantage of different teachers, we design a spatial-temporal co-teaching method for MVD. Specifically, we distill student models from both video teachers and image teachers by masked feature modeling. Extensive experimental results demonstrate that video transformers pretrained with spatial-temporal co-teaching outperform models distilled with a single teacher on a multitude of video datasets. Our MVD with vanilla ViT achieves state-of-the-art performance compared with previous methods on several challenging video downstream tasks. For example, with the ViT-Large model, our MVD achieves 86.4% and 76.7% Top-1 accuracy on Kinetics-400 and Something-Something-v2, outperforming VideoMAE by 1.2% and 2.4% respectively. When a larger ViT-Huge model is adopted, MVD achieves the state-of-the-art performance with **77.3%** Top-1 accuracy on Something-Something-v2. Code will be available at https://github.com/ruiwang2021/mvd.*

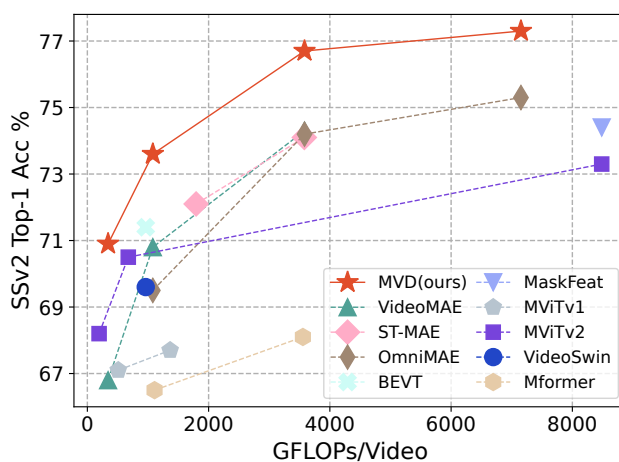
[†]Corresponding authors

Figure 1. **Comparisons of MVD with previous supervised or self-supervised methods on Something-Something v2.** Each line represents the corresponding model of different sizes.

1. Introduction

For self-supervised visual representation learning, recent masked image modeling (MIM) methods like MAE [31] and BEiT [2] achieve promising results with vision transformers [17] on various vision downstream tasks. Such a pretraining paradigm has also been adapted to the video domain and boosts video transformers by clear margins compared with supervised pretraining on several video downstream tasks. Representative masked video modeling (MVM) works include BEVT [63], VideoMAE [57] and ST-MAE [21].

Following MAE [31] and BEiT [2], existing masked video modeling methods [21, 57, 63] pretrain video transformers through reconstructing low-level features, *e.g.*, raw pixel values or low-level VQVAE tokens. However, using low-level features as reconstruction targets often incur much noise. And due to the high redundancy in video data, it is easy for masked video modeling to learn shortcuts, thus resulting in limited transfer performance on downstream

tasks. To alleviate this issue, masked video modeling [57] often uses larger masking ratios.

In this paper, we observe that much better performance on video downstream tasks can be achieved by conducting masked feature prediction by using the high-level features of pretrained MIM and MVM models as masked prediction targets. This can be viewed as two-stage masked video modeling, where MIM pretrained image models (*i.e.*, an image teacher) or MVM pretrained video models (*i.e.*, an video teacher) are obtained in the first stage, and they further act as teachers in the second stage for the student model via providing the high-level feature targets. Therefore, we call this method Masked Video Distillation (MVD).

More interestingly, we find that student models distilled with different teachers in MVD exhibit different properties on different video downstream tasks. Specifically, students distilled from the image teacher perform better on video tasks that mainly rely on spatial clues, while students distilled from the video teacher model perform better on the video downstream tasks where temporal dynamics are more necessary. We think during the pretraining process of masked video modeling in the first stage, video teachers have learned spatial-temporal context in their high-level features. Therefore, when employing such high-level representations as prediction targets of masked feature modeling, it will help encouraging the student model to learn stronger temporal dynamics. By analogy, image teachers provide high-level features as targets that include more spatial information, which can help the student model learn more spatially meaningful representations. We further analyze the feature targets provided by image teachers and video teachers, and calculate the cross-frame feature similarity. It shows that the features provided by the video teachers contain more temporal dynamics.

Motivated by the above observation, to leverage the advantages of video teachers and image teachers, we propose a simple yet effective spatial-temporal co-teaching strategy for MVD. In detail, the student model is designed to reconstruct the features coming from both the image teacher and video teacher with two different decoders, so as to learn stronger spatial representation and temporal dynamics at the same time. Experiments demonstrate that MVD with co-teaching from both the image teacher and the video teacher significantly outperforms MVD only using one single teacher on several challenging downstream tasks.

Despite the simplicity, our MVD is super effective and achieves very strong performance on multiple standard video recognition benchmarks. For example, on Kinectics-400 and Something-Something-v2 datasets, compared to the baseline without distillation, MVD with 400 epochs using a teacher model of the same size achieves **1.2%**, **2.8%** Top-1 accuracy gain on ViT-B. If a larger teacher model ViT-L is used, more significant performance gains (*i.e.*,

1.9%, **4.0%**) can be obtained. When ViT-Large is the target student model, our method can achieves **86.4%** and **76.7%** Top-1 accuracy on these two datasets, surpassing existing state-of-the-art method VideoMAE [57] by **1.2%** and **2.4%** respectively. When a larger ViT-Huge model is adopted, MVD achieves the state-of-the-art performance with **77.3%** Top-1 accuracy on Something-Something-v2.

Our contributions can be summarized as below:

- We find that using MIM pretrained image models and MVM pretrained video models as teachers to provide the high-level features for continued masked feature prediction can learn better video representation. And representations learned with image teachers and video teachers show different properties on different downstream video datasets.

- We propose masked video distillation together with a simple yet effective co-teaching strategy, which enjoys the synergy of image and video teachers.

- We demonstrate strong performance on multiple standard video recognition benchmarks, surpassing both the baseline without MVD and prior state-of-the-art methods by clear margins.

2. Related Work

Vision transformers for video understanding. For video understanding tasks, modeling the spatial-temporal information is the most important factor to consider in the architecture design. In the early works of video understanding, common video architectures, e.g., 3D CNNs [7, 20, 22, 58, 60] and 2D CNNs with temporal modules [12, 39, 53, 62, 73], are designed by extending existing 2D CNNs on the temporal dimension. Recently, Vision Transformers [13, 17, 41] achieve significant progress on several computer vision tasks. Some works [4, 18, 42, 69] also adapt vision transformers to the video domain and achieve superior performance compared to previous CNN-based architectures. For example, TimeSformer [4] and ViViT [1] study several variants of space-time factorization for the plain ViT. Some works [5, 47, 50] further explore how to reduce computational cost of the space-time attention. VideoSwin [42] and MViT [18, 38] study the hierarchical architecture and introduce an inductive locality bias into video transformers. Uniformer [36] and Video Mobile-Former [64] propose to integrate 3D CNNs and spatial-temporal attention mechanism for efficiency consideration. For convincing performance on the video understanding tasks, most video transformers require model weights pretrained on the large-scale image datasets. In this paper, we explore the self-supervised pretraining of video transformers and show pretraining strategy will significantly influence the downstream performance, which is orthogonal to the transformer architecture design.

Self-supervised video representation learning. The early works [3, 45, 66, 70] of self-supervised video representation learning focus on designing the pretext tasks based on the temporal structure of videos. More recently, contrastive learning [8, 32, 37] that forces different views of the same image sample to be closer in the feature space while pushing the views of different images farther becomes a new paradigm of representation learning, and some works [11, 23, 24, 29, 30, 46, 49] design the contrastive learning methods on the video domain by exploring effective ways of spatial-temporal augmentations. However, as the learning supervision based on contrastive learning is applied on global representation, it cannot well model the local relationship or learn fine-grained local representation.

Masked visual modeling. Masked language modeling [10, 40] has been one of the dominant pretraining methods of language transformers. With the success of vision transformers, masked visual modeling [2] has been introduced to self-supervised visual pretraining and demonstrates to be also helpful to multimodal visual-language learning [16, 74]. Following BERT [10] pretraining, BEiT [2] and PeCo [15] pretrain ViT by predicting the discrete visual tokens of masked patches, which are encoded by a pretrained VQ-VAE. MAE [31] proposes an asymmetric encoder-decoder framework for the reconstruction of pixels. Sim-MIM [68] and MaskFeat [67] propose to recover low-level features of masked patches like pixels or HOG features with hierarchical ViT. In contrast, some works [9, 14, 75] adopt an exponential moving average of the student model as the online teacher model, which makes the target features boot-strapped during training. In the video domain, some pioneering works [21, 55, 57, 63] extend masked image modeling to masked video modeling. BEVT [63] proposes a two-stream pretraining joint pretraining framework by predicting the discrete tokens with both image transformer and video transformer. VideoMAE [57] and ST-MAE [21] follow MAE and reconstruct the pixels of masked video patches with an extremely high masking ratio. Unlike most previous works of masked video modeling, our MVD focuses on masked feature modeling with high-level features as targets, and finds that student models using image and video teacher models will have different properties and complement each other.

Knowledge distillation. Knowledge distillation [26, 33, 48] aims to transfer the knowledge of the teacher model to the student model by adopting the output of the teacher model as the target for training the student model. Typical knowledge distillation works [33, 51, 52] mainly focus on supervised learning, e.g., image classification. The idea of knowledge distillation has also been leveraged in many self-supervised methods by learning from momentum encoders [6, 56] or different architectures [24]. Recently, self-supervised knowledge distillation [19, 71, 72] has also been studied to learn representations from self-supervised pretrained models. In this paper, we present the first attempt that uses the masked image modeling pretrained image and video model as the masked feature prediction target in the video domain. It shows self-supervised MIM pretrained model can further boostrap the mask video pretraining and bring significant performance gain.

3. Method

While masked video modeling has demonstrated promising performance for self-supervised learning, most existing approaches reconstruct relatively low-level information in the forms of raw pixels [57], low-level features like HOG [67] and VQVAE tokens [63]. In this paper, instead of reconstructing low-level information, we conduct masked video modeling at the feature-level. This is achieved by a two-stage framework, MVD, optimized to predict high-level features that derived from off-the-shelf MIM pretrained image models [31] and MVM pretrained video models [57] which are readily available. Below, we first given an overview of the masked feature modeling paradigm in Sec. 3.1 and then we introduce our proposed MVD in Sec. 3.2. Finally, we present the architectural design of MVD in Sec. 3.3.

3.1. The Paradigm of Masked Feature Modeling

The core of masked feature modeling is to train models to predict features of masked input regions. In this paper, we follow the decoupled encoder-decoder transformer architecture in MAE [31] due to its effectiveness and simplicity. An input X (image $X_{img} \in \mathbb{R}^{H \times W \times 3}$ or video $X_{vid} \in \mathbb{R}^{T \times H \times W \times 3}$) is partitioned into several non-overlapping patches, and then each patch is mapped to a visual token with a linear projection layer. Before feeding tokens to the transformer encoder f, a subset of tokens is masked and dropped from the token sequence. To reconstruct the information of masked tokens, the token sequence consisting of the visible tokens from the encoder and learnable mask tokens are input to a shallow transformer decoder g:

$$Y = g(\texttt{concat}(f(X_{vis}), T_m)), \tag{1}$$

where X_{vis} denotes the visible input tokens, and T_m denotes the mask tokens. The subset of output tokens from decoder corresponded to input mask tokens contains reconstructed information of masked tokens. The reconstruction target for each masked patch $X(p)$ is represented as a patch feature $h(X(p))$. Here, h is represents a function for generating the target features, *e.g.* h produces low-level RGB values of pixels in the patch in [21, 57]. Then, to train the encoder and the decoder, a loss function that measures the distance D between the ground-truth features of masked patches and reconstructed ones is defined as:
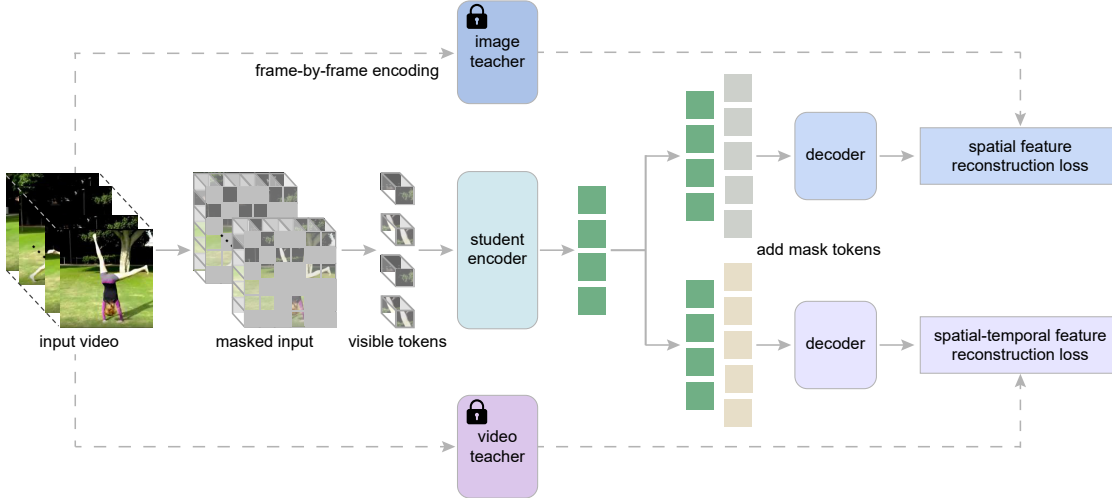
Figure 2. **An overview of MVD framework.** Firstly the image teacher is pretrained by masked image modeling and the video teacher is pretrained by masked video modeling. Then the student model is trained from scratch to predict target high-level features encoded by the image teacher and the video teacher. The teacher models are fixed in the distillation stage.

$$L_{mfm}(h) = \frac{1}{|M|} \sum_{p \in M} D(Y(p), h(X(p))) \qquad (2)$$

where p is the token index and M is the set of masked tokens. For pixel regression in MAE [31] and Video-MAE [57], the L2 distance is used as the distance metric.

3.2. Masked Video Distillation

In this paper, we propose Masked Video Distillation (MVD), which performs masked feature modeling on videos using high-level features as opposed to low-level pixels. In particular, we simply use outputs generated by off-the-shelf self-supervised pretrained image or video models, which are readily available, as reconstruction targets. These high-level features, serving as targets of the mask & prediction tasks, are encoded by teacher models pretrained by masked visual modeling like MAE or Video-MAE. For video representation learning, the reconstruction targets can take the form of spatial features encoded by image teacher models, or spatial-temporal features encoded by video teacher models. More specifically, the image teachers is pretrained by masked image modeling, while the video teacher is pretrained with masked video modeling, both of which aim at reconstructing raw pixels. Once trained, we use the image encoder h_{img} to generate the spatial targets, and the pretrained video transformer encoder h_{vid} to generate spatial-temporal targets. The loss function of MVD with the image teacher and video teacher can be denoted by $L_{mfm}(h_{img})$ and $L_{mfm}(h_{vid})$, respectively

Spatial-temporal Co-teaching. When performing MVD with a single teacher, we observe that students distilled from different teachers learn different video representations and perform well on different kinds of downstream video

tasks. To improve the performance of MVD on different downstream video tasks, we propose spatial-temporal co-teaching that explores information from both image and video teachers such that the student model can handle videos of different types better. For instance, videos with fastly changing human actions require more temporal information while spatial clues might be sufficient for relatively static videos. To this end, MVD is trained to predict target high-level features produced by the image teacher and the video teacher at the same time. This is achieved by using two separated decoders to reconstruct different target features. The final loss of MVD with spatial-temporal co-teaching is:

$$L_{mvd} = \lambda_1 L_{mfm}(h_{img}) + \lambda_2 L_{mfm}(h_{vid}) \qquad (3)$$

where λ_1 and λ_2 denote the hyper-parameters that balance the weights of the image teacher and the video teacher. The pseudo code of MVD is shown in Algorithm 1.

3.3. Architectural Design

Encoder. For MVD, the vanilla transformer backbone is used as the encoder. For a video input $X_{vid} \in \mathbb{R}^{T \times H \times W \times 3}$, we adopt 3D patch embedding with a patch size of $2 \times 16 \times 16$. After patch partitioning and linear embedding, we obtain $T/2 \times H/16 \times W/16$ tokens. For the masked feature modeling, the tokens are masked with a high masking ratio and the remaining tokens are fed into the transformer layers. For finetuning on downstream tasks, we input all tokens to the subsequent layers. In each layer, joint spatial-temporal self-attention is applied on the whole input token sequence.

Mask strategy. For MVD, we follow [57] and adopt tube masking for masked feature modeling. First a 2D random

Algorithm 1 Pseudocode of MVD in PyTorch style.

```
# f: student encoder
# g_img: decoder for reconstructing spatial features
# g_vid: decoder for reconstructing spatial-temporal
    features
# t_m: learnable mask tokens
# h_img: image teacher model
# h_vid: video teacher model

for x, m in loader: # x: video data, m: mask
    x_pe = patch_emb(x) # patch embedding of input
    x_vis = mask_select(x_pe, 1 - m) # masking tokens
    q_vis = f(x_vis) # visible local patch features

    # reconstruction of target features
    p_img = g_img(concat(q_vis, t_m))
    p_vid = g_vid(concat(q_vis, t_m))

    # compute target features with teacher models
    k_img = h_img(x) # target spatial features
    k_vid = h_vid(x) # target spatial-temporal features

    # compute reconstruction loss
    loss_img = smooth_L1_loss(p_img ⊙ m, k_img ⊙ m)
    loss_vid = smooth_L1_loss(p_vid ⊙ m, k_vid ⊙ m)

    loss = λ₁ * loss_img + λ₂ * loss_vid
    loss.backward()
    optimizer.step() # optimizer update
```

mask is generated and then extended along the temporal dimension. Therefore, the spatial mask on each time slice is the same, which prevents information leakage between frames. Tube masking with a high masking ratio (*e.g.*, 90%) encourages the video transformer to model the high-level semantics during pretraining.

Decoder. For MVD, shallow decoders consist of vanilla transformer layers and a linear projection layer. The transformer layers in decoders are the same as those in the encoder. Since spatial-temporal co-teaching introduces two different reconstruction targets for masked feature modeling, two separated decoders that share the same architecture but contain different weights are placed on the top of the encoder. Learnable masked tokens corresponded to masked patches are concatenated with visible tokens from the encoder before fed into the decoder. After jointly modeling the spatial-temporal relationships, the output tokens of transformer layers are mapped to final predictions by the linear projection layer.

Reconstruction targets. For generating spatial-temporal target features, the video teacher, which shares the same architecture as the student model, is pretrained by a Video-MAE [57] manner on the video dataset. For obtaining spatial targets, we adopt the vanilla image ViT pretrained by masked image modeling [31] on the image dataset (*e.g.*, ImageNet-1K). It is worth noting that one 3D patch (with size of $2 \times 16 \times 16$) for the video transformer corresponds to two 2D patches (with size of 16×16) for the image transformer. Following [21], we predict the spatial features of a single time slice (that is the front 2D patch), which reduce the prediction layer's size.

4. Experiments

In this section, we first introduce the experimental setup in Sec. 4.1, and then present the main results in Sec. 4.2, followed by an extensive analysis to verify the effectiveness of different components in Sec. 4.3.

4.1. Experimental Setup

Dataset. We pretrain the vanilla ViT with MVD on Kinetics-400 by default, and evaluate the learned model on four video recognition downstream tasks: (a) Kinetics-400 (K400) [7], which consists of $\sim 240K$ training videos and $\sim 20K$ validation videos with an average duration of 10 seconds. All video clips are labeled into 400 classes. (b) Something-Something V2 (SSv2) [27], which contains $\sim 160K$ videos for training and $\sim 20K$ videos for validation. The videos in SSv2 with an average duration of 4 seconds are labeled into 174 motion-centric categories. (c) UCF-101 [54] is a relatively small dataset, consisting of $\sim 9.5K$ training videos and $\sim 3.5K$ validation videos. (d) HMDB51 [35] is also a small video dataset that contains around 3.5K/1.5K train/val videos. We also transfer pretrained models to a challenging spatial-temporal action detection dataset AVA [28]. The results of AVA are included in supplementary materials.

Implementation details. Our MVD is performed on vanilla ViTs with different capacities (*i.e.*, ViT-S, ViT-B, ViT-L). By default, image teacher models are pretrained on ImageNet-1K for 1600 epochs and video teachers are pretrained on K400 for 1600 epochs. We follow the training strategy in MAE [31] and VideoMAE [21] for image teachers and video teachers respectively. In the distillation stage, student models are first pretrained from scratch on K400 for 400 epochs unless mentioned otherwise. The resulting models are then finetuned on downstream video tasks. The video clip length is 16 for both pretraining and finetuning. We adopt AdamW optimizer [44] and Smooth L1 loss for the optimization of student models. More details are included in supplementary materials.

4.2. Main Results

Students distilled from different teachers. Unlike masked image modeling, masked feature modeling on video data has more choices on reconstruction targets. Besides spatial features, to include temporal dynamics in the reconstruction target, we can also adopt spatial-temporal features encoded by pretrained video models. In Table 1, we compare students distilled by the image teacher and the video teacher on K400, a downstream task that mainly relies on spatial clues, and SSv2, a temporally-heavy downstream task. Our observations are as follow: (a) Masked feature modeling with high-level features as targets achieves convincing performance on downstream video tasks, and outperforms Video-

student	teachers		K400 top-1	SSv2 top-1
	image	video		
ViT-S	✓	✗	80.4	69.4
	✗	✓	80.1	70.0
	✓	✓	**80.6**	**70.7**
ViT-B	✓	✗	82.3	71.4
	✗	✓	82.1	71.8
	✓	✓	**82.7**	**72.5**

Table 1. **Co-teaching with the image teacher and the video teacher outperforms distillation with a single teacher in MVD.** We adopt ViT-B as the architecture of teachers and distill students for 400 epochs here.

student	teacher	K400 top-1		SSv2 top-1	
		ViMAE	MVD	ViMAE	MVD
ViT-S	ViT-B	79.0	80.6 ↑**1.6**	66.4	70.7 ↑**4.3**
ViT-S	ViT-L	79.0	81.0 ↑**2.0**	66.4	70.9 ↑**4.5**
ViT-B	ViT-B	81.5	82.7 ↑**1.2**	69.7	72.5 ↑**2.8**
ViT-B	ViT-L	81.5	83.4 ↑**1.9**	69.7	73.7 ↑**4.0**
ViT-L	ViT-L	85.2	86.0 ↑**0.8**	74.0	76.1 ↑**2.1**

Table 2. **MVD achieves significant improvement compared with VideoMAE across different model scales.** MVD is pretrained on K400 for 400 epochs, while VideoMAE is pretrained on K400 for 1600 epochs here.

MAE baseline significantly (compared with the baseline results in Table 2). In particular, with both image and video teachers using a ViT-S as the backbone, MVD achieves consistent gains over VideoMAE on both K400 (80.6% *vs*. 79.0%) and SSV2 (70.7% *vs*. 66.4%). (b) Students distilled from the image teacher achieve higher top-1 accuracy on K400, while students distilled from the video teacher perform better on SSv2. For example, ViT-S achieves an accuracy of 80.4% and 69.4% using an image teacher on K400 and SSv2 respectively. With a video teacher, on the other hand, ViT-S offers a top-1 accuracy of 80.1% and 70.0% respectively. As it has been shown that videos in K400 less sensitive to temporal modeling compared to SSv2, the results demonstrate students learn stronger spatial representation from the image teacher while the video teacher transfer more knowledge about temporal dynamics to students.

Co-teaching outperforms distilling with a single teacher. To improve the performance on different kinds of downstream video tasks, we introduce spatial-temporal co-teaching in MVD, which trains the model to predict spatial features and spatial-temporal features of masked patches in a decoupled way. The results in Table 1 indicate that students distilled with spatial-temporal co-teaching outperforms students distilled from either single teacher on both spatially-heavy task and temporally-heavy task.

MVD outperforms VideoMAE baseline significantly. In Table 2, MVD with spatial-temporal co-teaching is compared with VideoMAE pretrained on K400. When the size of teacher models is the same as that of students, MVD out-

method	extra data	top-1	top-5	GFLOPs	Param
supervised					
NL I3D R101 [65]	-	77.3	93.3	359×30	62
ip-CSN-152 [59]	-	77.8	92.8	109×30	33
SlowFast NL [22]	-	79.8	93.9	234×30	60
X3D-XL [20]	-	79.1	93.9	48×30	11
MViTv1-B [18]	-	80.2	94.4	170×5	37
VideoSwin-B [42]	IN-1K	80.6	94.6	282×12	88
Uniformer-B [36]	IN-1K	83.0	95.4	259×12	50
TimeSformer [4]	IN-21K	80.7	94.7	2380×3	121
Mformer-B [47]	IN-21K	79.7	94.2	370×30	109
Mformer-L [47]	IN-21K	80.2	94.8	1185×30	382
ViViT-L FE [1]	IN-21K	81.7	93.8	3980×3	N/A
VideoSwin-L [42]	IN-21K	83.1	95.9	604×12	197
self-supervised					
VIMPAC ViT-L [55]	HowTo100M	77.4	N/A	N/A×30	307
BEVT Swin-B [63]	IN-1K	81.1	N/A	282×12	88
MaskFeat MViT-S [67]	-	82.2	95.1	71×10	36
VideoMAE ViT-S [57]	-	79.0	93.8	57×15	22
VideoMAE ViT-B [57]	-	81.5	95.1	180×15	87
VideoMAE ViT-L [57]	-	85.2	96.8	597×15	305
VideoMAE ViT-H [57]	-	86.6	97.1	1192×15	633
ST-MAE ViT-B [21]	-	81.3	94.9	180×21	87
ST-MAE ViT-L [21]	-	84.8	96.2	598×21	304
ST-MAE ViT-H [21]	-	85.1	96.6	1193×21	632
OmniMAE ViT-B [25]	IN-1K	80.8	N/A	180×15	87
OmniMAE ViT-L [25]	IN1K+SSv2	84.0	N/A	597×15	305
OmniMAE ViT-H [25]	IN1K+SSv2	84.8	N/A	1192×15	633
MVD-S (Teacher-B)	IN-1K	80.6	94.7	57×15	22
MVD-S (Teacher-L)	IN-1K	81.0	94.8	57×15	22
MVD-B (Teacher-B)	IN-1K	82.7	95.4	180×15	87
MVD-B (Teacher-L)	IN-1K	83.4	95.8	180×15	87
MVD-L (Teacher-L)	IN-1K	86.0	96.9	597×15	305
MVD-L (Teacher-L) †	IN-1K	**86.4**	**97.0**	597×15	305
MVD-H (Teacher-H) †	IN-1K	**87.3**	**97.4**	1192×15	633

Table 3. **Comparison with previous works on Kinetics-400.** "MVD-S (Teacher-B)" denotes that the student is ViT-S and the teacher is ViT-B. † denotes that the student is distilled for 800 epochs instead of 400 epochs. The inference cost is reported with the cost of a single view × the number of views. Note that IN-1K is only used for training image teachers in MVD.

performs VideoMAE by a clear margin on both K400 and SSv2. Larger models as teachers can further boost the performance of MVD. It is worth mentioning that not only is our MVD particularly effective for relatively small models, but also it improves the performance of large vision models like ViT-L. For example, with a ViT-L model as the student, MVD achieves 86.0% and 76.1% on K400 and SSv2, surpassing VideoMAE by 0.8% and 2.1%, respectively.

Comparison with state-of-the-art. We compare MVD with prior studies on four video recognition tasks. Results on K400 are shown in Table 3. Our MVD outperforms previous self-supervised methods with similar or less computational cost. Even compared with video transform-

method	extra data	top-1	GFLOPs	Param
supervised				
SlowFast R101 [22]	K400	63.1	106×3	53
TSM-RGB R50 [39]	IN-1K	63.3	62×6	24
TAM R50 [43]	IN-1K	66.0	99×6	51
TDN R101 [61]	IN-1K	69.6	198×3	88
MViTv1-B [18]	-	67.7	455×3	37
MViTv2-B [38]	K400	70.5	225×3	51
Uniformer-B [36]	K400	71.2	259×3	50
TimeSformer-HR [4]	IN-21K	62.5	1703×3	121
ViViT-L FE [1]	IN21K+K400	65.9	995×12	N/A
Mformer-B [47]	IN21K+K400	66.5	370×3	109
Mformer-L [47]	IN21K+K400	68.1	1185×3	382
VideoSwin-B [42]	IN21K+K400	69.6	321×3	88
MViTv2-L [38]	IN21K+K400	73.3	2828×3	213
self-supervised				
VIMPAC ViT-L [55]	HowTo100M	68.1	N/A×30	307
BEVT Swin-B [63]	IN1K+K400	71.4	321×3	88
MaskFeat MViT-L [67]	K400	74.4	2828×3	218
VideoMAE ViT-S [57]	K400	66.4	57×6	22
VideoMAE ViT-S [57]	-	66.8	57×6	22
VideoMAE ViT-B [57]	K400	69.7	180×6	87
VideoMAE ViT-B [57]	-	70.8	180×6	87
VideoMAE ViT-L [57]	K400	74.0	597×6	305
VideoMAE ViT-L [57]	-	74.3	597×6	305
ST-MAE ViT-L [21]	K400	72.1	598×3	304
ST-MAE ViT-H [21]	K400	74.1	1193×3	632
OmniMAE ViT-B [25]	IN1K+K400	69.0	180×6	87
OmniMAE ViT-L [25]	IN-1K	74.2	597×6	305
OmniMAE ViT-H [25]	IN-1K	75.3	1192×6	632
MVD-S (Teacher-B)	IN1K+K400	70.7	57×6	22
MVD-S (Teacher-L)	IN1K+K400	70.9	57×6	22
MVD-B (Teacher-B)	IN1K+K400	72.5	180×6	87
MVD-B (Teacher-L)	IN1K+K400	73.7	180×6	87
MVD-L (Teacher-L)	IN1K+K400	76.1	597×6	305
MVD-L (Teacher-L) †	IN1K+K400	**76.7**	597×6	305
MVD-H (Teacher-H) †	IN1K+K400	**77.3**	1192×6	633

Table 4. **Comparison with previous works on Something-Something V2..** † denotes that the student is distilled for 800 epochs.

ers pretrained on ImageNet-21K, MVD achieves superior performance. Particularly, MVD-H achieves an accuracy of 87.3% on K400, outperforming previous top-performing methods by clear margins. Table 4 presents comparisons to the state-of-the-art methods on SSv2. We observe that for downstream tasks depending on temporal relationship modeling, self-supervised methods based on MVM achieve better performance in comparison with supervised methods (*cf*. results in the middle group of Table 4 *vs*. results in the top group). Once again, our MVD, producing an accuracy of 76.1% with a large model, beats both supervised meth-

method	extra data	Param	UCF101	HMDB51
VideoMoCo R2+1D [46]	K400	15	78.7	49.2
MemDPC R2D3D [30]	K400	32	86.1	54.5
Vi^{2}CLR S3D [11]	K400	9	89.1	55.7
CORP Slow-R50 [34]	K400	32	93.5	68.0
CVRL Slow-R50 [49]	K400	32	92.9	67.9
CVRL Slow-R152 [49]	K600	328	94.4	70.6
ρBYOL Slow-R50 [23]	K400	32	94.2	72.1
VIMPAC ViT-L [55]	HowTo100M	307	92.7	65.9
VideoMAE ViT-B [57]	K400	87	96.1	73.3
MVD-B (Teacher-B)	IN-1K+K400	87	**97.0**	**76.4**
MVD-B (Teacher-L)	IN-1K+K400	87	**97.5**	**79.7**

Table 5. **Comparison with previous state-of-the-art methods on UCF101 and HMDB51**.
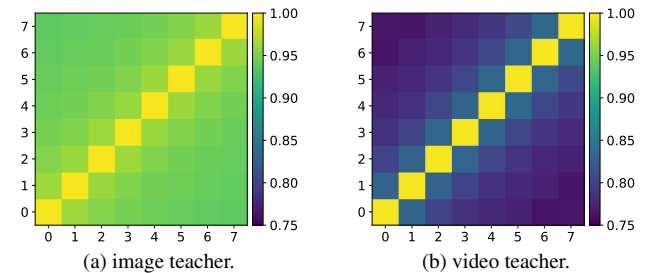

(a) image teacher.   (b) video teacher.

Figure 3. **Feature similarity across different frames for different teacher models.** Similarity matrices are computed on the Kinetics-400 validation set.

ods and self-supervised methods by clear margins. With more training epochs (*i.e.*, 800 epochs), MVD with ViT-L achieves more significant performance gains (*i.e.*, 1.2%, 2.4% on K400 and SSv2) compared with VideoMAE. When a huge model is used, the performance can still be boosted and MVD achieves 77.3% top-1 accuracy on SSv2.

We also evaluate the transfer learning ability of MVD on two relatively small datasets, UCF101 and HMDB51. As shown in Table 5, MVD with ViT-B obtains higher accuracy compared with prior works based on well-designed pretext tasks, contrastive learning and masked video modeling methods. Especially compared to the original Video-MAE [57] ViT-B teacher model, we achieve 0.9% and 3.1% higher points on these two datasets respectively. Additionally, when teachers with larger size are adopted, MVD achieves stronger transfer learning performance.

4.3. Analysis and Discussion

In this section, we provide an in-depth analysis of the effectiveness of different components in MVD.

Analysis of features encoded by different teachers. The properties of target features generated by different teachers may influence the performance of students on different downstream tasks. To quantify the temporal dynamics that teacher models capture from the input video, we study the similarity between feature maps across different frames of each input video clip via the cosine similarity. As similarity

method	epoch	time(h)	K400 top-1
VideoMAE	800	107	81.0
VideoMAE	1600	214	81.5
MVD	(800+) 400	(107+) 57	**81.9**

Table 6. **Training time comparison between MVD and Video-MAE on 32 NVIDIA V100 GPUs.** The training time of teacher models is considered. we only adopt a video teacher pretrained for 800 epochs in MVD here. ViT-B is used here.

teachers	reconstruct pixel	SSv2 top-1
image model	✗	68.7
image model	✓	67.9 ↓**0.8**
image+video models	✗	70.1
image+video models	✓	69.0 ↓**1.1**

Table 7. **Ablations on reconstructing low-level features in masked video distillation.** We distill ViT-S students with ViT-B teachers for 300 epochs here.

teacher	init	epoch	top-1	
			K400	SSv2
momentum encoder	✗	800	80.5	70.4
momentum encoder	✓	800	81.8	70.8
fixed image model	✗	400	82.3	71.4
fixed video model	✗	400	82.1	71.8
fixed co-teaching	✗	400	82.7	72.5

Table 8. **Comparison with bootstrapped teachers and students initialized with IN1K-pretrained models.** ViT-B is used here.

matrices shown in Figure 3, for image teachers, the feature maps of different frames are almost the same. However, for video teachers, the features of different frames have larger difference. This indicates that video teachers capture more temporal difference. Therefore, students distilled from video teachers can learn stronger temporal dynamics and perform better on temporally-heavy downstream tasks.

Training time comparison. We study whether MVD is able to achieve better balance of accuracy and efficiency than VideoMAE. For fair comparisons, the training time of teacher models is also counted in the total training time of MVD. Results are shown in Table 6. We see that MVD can achieve better accuracy (*i.e.*, 81.9%) by a total of 164 hours of training which is 50 hours less than VideoMAE (producing an accuracy of 81.5%) trained for 1600 epochs.

Reconstruction signals in MVD. In MVD, we first pre-train teacher models by recovering the pixels of masked patches in a MAE manner, then adopt features produced by teachers as the targets of masked feature modeling. In Table 7, we study whether to include an additional decoder branch for reconstructing pixels of masked patches in the distillation stage. The experimental results show that for both MVD with a single teacher and MVD with spatial-temporal co-teaching, the reconstruction of low-level feature targets degrades the performance for downstream tasks. Therefore, we only reconstruct high-level features of masked patches in the distillation stage of MVD.

distillation method	K400 top-1	SSv2 top-1
per-token distillation	80.9	70.5
masked reconstruction	82.1	71.8

Table 9. **Comparison with feature distillation without masked reconstruction.** We distill ViT-B from the video teacher here.

Comparison with bootstrapped teachers. Some recent works of image representation learning [9, 14, 16, 75] adopt features of a momentum encoder as the targets of MIM, while we use frozen teacher models in MVD. In Table 8, we compare fixed teachers with bootstrapped teachers that are updated by an exponential moving average of the online encoder during pretraining. According to a MIM method [14], two strong baselines of bootstrapped teachers are built for video representation learning: (a) The student model is trained from scratch with masked feature modeling, and the target features are generated by a momentum encoder. (b) The framework of bootstrapped teachers is first pretrained on IN-1K for 800 epochs, then the pretrained weights are adopted to initialize the video pretraining. As shown in Table 8, MVD with the frozen teacher beats the method with a bootstrapped teacher on the downstream video tasks, even if only a single teacher is utilized.

Comparison with feature distillation. In the previous work of self-supervised feature distillation [19, 26, 72], the distillation loss is directly computed upon the full feature maps between teachers and students. Accordingly, we use a baseline method named per-token distillation. Specifically, the output features of students are projected by a MLP, and then forced to mimicking the teacher's feature at each token with a Smooth L1 loss. As shown in Table 9, masked feature reconstruction in our MVD outperforms per-token distillation on both K400 and SSv2.

5. Conclusion

In this paper, we study masked video distillation upon MIM pretrained image or video transformers. We have three interesting findings: 1) Using MIM pretrained image transformers or MVM pretrained video transformers as teachers supervise masked feature prediction can significantly boost the finetuning performance on video downstream tasks; 2) The representation distilled with image and video teachers will have different properties, i.e., image teachers will benefit spatial-heavy video tasks more while video teachers benefit temporal-heavy video tasks more; 3) Combining image and video teachers will enjoy the synergy and thus produce higher performance. Even though the proposed masked video distillation seems very straightforward, we hope such interesting findings can motivate more thinking about masked video pretraining.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2, 6, 7

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 1, 3

[3] 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 *CVPR*, 2020. 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2, 6, 7

[5] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *NeurIPS*, 2021. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[9] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *ECCV*, 2022. 3, 8

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[11] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *ICCV*, 2021. 3, 7

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022. 2

[14] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *ECCV*, 2022. 3, 8

[15] 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

[16] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 3, 8

[17] 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. In *ICLR*, 2021. 1, 2

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2, 6, 7

[19] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 3, 8

[20] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2, 6

[21] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1, 3, 5, 6, 7

[22] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2, 6, 7

[23] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 2021. 3, 7

[24] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attention via transformers in self-supervised visual representation learning. *NeurIPS*, 2021. 3

[25] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 6, 7

[26] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 2021. 3, 8

[27] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 5

[28] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 5

[29] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *CVPR*, 2022. 3

of vision transformers. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 3, 7

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 3, 4, 5

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[33] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[34] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *ICCV*, 2021. 7

[35] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 5

[36] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *ICLR*, 2022. 2, 6, 7

[37] Suichan Li, Dongdong Chen, Yinpeng Chen, Lu Yuan, Lei Zhang, Qi Chu, Bin Liu, and Nenghai Yu. Improve unsupervised pretraining for few-label transfer. In *ICCV*, 2021. 3

[38] 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 *CVPR*, 2022. 2, 7

[39] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2, 7

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2, 6, 7

[43] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *CVPR*, 2021. 7

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 5

[45] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pages 527–544. Springer, 2016. 3

[46] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *CVPR*, 2021. 3, 7

[47] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 2, 6, 7

[48] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *ICML*, 2019. 3

[49] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 3, 7

[50] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *NeurIPS*, 2021. 2

[51] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. *arXiv preprint arXiv:2009.08453*, 2020. 3

[52] Zhiqiang Shen and Eric Xing. A fast knowledge distillation framework for visual recognition. *arXiv preprint arXiv:2112.01528*, 2021. 3

[53] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 2

[54] 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. 5

[55] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021. 3, 6, 7

[56] Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. Isd: Self-supervised learning by iterative similarity distillation. In *ICCV*, 2021. 3

[57] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1, 2, 3, 4, 5, 6, 7

[58] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: Generic features for video analysis. In *ICCV*, 2015. 2

[59] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 6

[60] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 2

[61] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 2021. 7

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 2018. 2

[63] 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 *CVPR*, 2022. 1, 3, 6, 7

[64] Rui Wang, Zuxuan Wu, Dongdong Chen, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Luowei Zhou, Lu Yuan, and Yu-Gang Jiang. Video mobile-former: Video recognition with efficient global spatial-temporal modeling. *arXiv preprint arXiv:2208.12257*, 2022. 2

[65] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 6

[66] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 3

[67] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 3, 6, 7

[68] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 3

[69] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. *arXiv preprint arXiv:2211.13222*, 2022. 2

[70] 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*, pages 10334–10343, 2019. 3

[71] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, 2020. 3

[72] Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *ICLR*, 2021. 3, 8

[73] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 2

[74] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022. 3

[75] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 8