Asymmetric Masked Distillation for Pre-Training Small Foundation Models

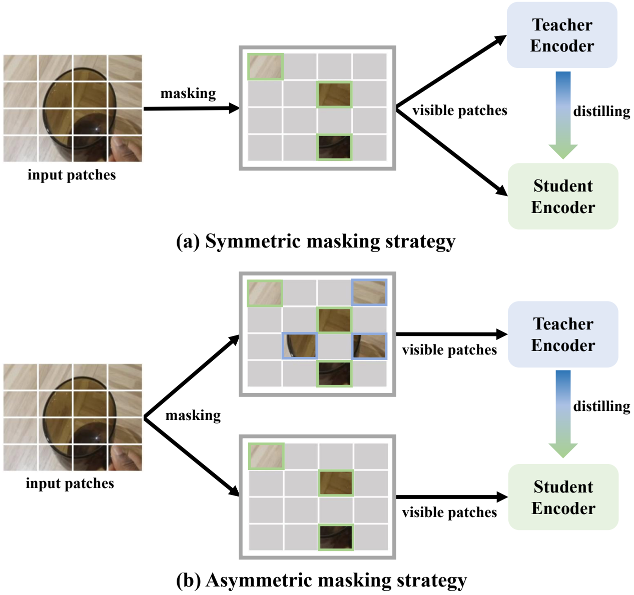
Zhiyu Zhao[1,2] Bingkun Huang[1,2] Sen Xing[2] Gangshan Wu[1] Yu Qiao[2] Limin Wang[1,2,✉]
[1] State Key Laboratory for Novel Software Technology, Nanjing University [2] Shanghai AI Lab

Abstract

Self-supervised foundation models have shown great potential in computer vision thanks to the pre-training paradigm of masked autoencoding. Scale is a primary factor influencing the performance of these foundation models. However, these large foundation models often result in high computational cost. This paper focuses on pre-training relatively small vision transformer models that could be efficiently adapted to downstream tasks. Specifically, taking inspiration from knowledge distillation in model compression, we propose a new asymmetric masked distillation (AMD) framework for pre-training relatively small models with autoencoding. The core of AMD is to devise an asymmetric masking strategy, where the teacher model is enabled to see more context information with a lower masking ratio, while the student model is still equipped with a high masking ratio. We design customized multi-layer feature alignment between the teacher encoder and student encoder to regularize the pre-training of student MAE. To demonstrate the effectiveness and versatility of AMD, we apply it to both ImageMAE and VideoMAE for pre-training relatively small ViT models. AMD achieved 84.6% classification accuracy on IN1K using the ViT-B model. And AMD achieves 73.3% classification accuracy using the ViT-B model on the Something-in-Something V2 dataset, a 3.7% improvement over the original ViT-B model from VideoMAE. We also transfer AMD pre-trained models to downstream tasks and obtain consistent performance improvement over the original masked autoencoding. The code and models are available at https://github.com/MCG-NJU/AMD.

1. Introduction

In recent years, self-supervised learning (SSL) [6, 10, 12, 24, 29, 62] has witnessed great success and outperformed its supervised counterparts. With the success in masked language modeling (MLM) [15], masked image modeling has become popular in computer vision for self-supervised representation learning. For example, BeiT [4], SimMIM [68], and MAE [27] are proposed to image masked pre-training, and MaskFeat [64], VideoMAE [51], MAE-ST [22], VideoMAEv2 [57], and MGMAE [33] are developed for video masked representation learning. This simple pipeline of masking and reconstruction has shown excellent performance on downstream tasks such as image classification, object detection, semantic segmentation, and action recognition. However, some issues still remain for the masked autoencoding framework. First, the encoder often operates on a small portion of visible tokens with a high masking ratio (*e.g.*, 75% for image and 90% for video). This high masking could increase the difficulty of the pre-training tasks and might encourage the encoder to capture more useful high-level information for reconstruction. We argue that this high masking ratio might also lose some important and detailed structure information, leading to the pre-trained model capturing incomplete and biased visual information. Second, the masked autoencoding often requires the ViT backbones of high capacities (*e.g.*, ViT-Large and ViT-Huge) to unleash



(a) Symmetric masking strategy

(b) Asymmetric masking strategy

Figure 1. **Comparison of symmetric and asymmetric masking strategy.** The asymmetric masking strategy allows the teacher to acquire more contextual information than the students.

✉: Corresponding author (lmwang@nju.edu.cn).

the power of the masked pre-training. These large ViT models take high computational cost and memory consumption during fine-tuning on the variety of downstream tasks. This high cost is particularly severe for video input as video transformer takes multiple frames as inputs. We think that we should pay more attention to the relatively small ViT models in masked autoencoding, which could have higher efficiency and more application potential in downstream tasks.

To overcome the above issues in masked autoencoding, we resort to the general paradigm for knowledge distillation [31] in model compression. It provides an effective *teacher* to *student* training framework to transfer the dark knowledge in powerful models to the lightweight student model. We extend this idea to the masked autoencoding paradigm to build a more efficient and effective pre-trained model, which could be applied to a variety of downstream tasks. Wei *et al*. [65] applies the feature alignment to distill the unsupervised pre-trained model, but they found that this method yielded little benefit for the MAE pre-trained model, as it already had diverse attention heads. Recently, Bai *et al*. [3] proposed a scheme for MAE distillation (DMAE) where feature distillation is performed alongside pre-training for the reconstruction task. DMAE allows the student and the teacher to receive the symmetric unmasked patches so that features can be aligned directly and the computational complexity of the teacher model can be reduced. But the same masking for both teacher and student limits the teacher from gathering more context information from inputs, and still faces the information loss risk as mentioned above.

Based on the above analysis, we propose an asymmetric masked distillation structure for MAE pre-training and the goal is to obtain a small but robust pre-trained MAE model. The masking ratio of the student remains at its default setting, while the masking ratio of the teacher is relatively reduced. And the unmasked patches of the student are a subset of those of the teacher, as in Figure 1. This asymmetric masked distillation maintains the difficulty of the reconstruction task for the students during the pre-training process and also allows the teacher to receive more context information that can be transferred to the student. However, it is not reasonable to significantly reduce the masking ratio of the teacher, as the teacher takes up a high amount of computational resources. Hence we have made a proper compromise between the masking ratio of the teacher and the computational cost. In our asymmetric structure, the visible patches are divided into two types, one is visible to both student and teacher, and the other is visible only to the teacher. To align these two types of features, a serial alignment strategy is applied.

With this efficient design of asymmetric masked distillation, we were able to achieve a performance of 73.3% on SSV2 based on VideoMAE, leaving a gap of only 1% with the larger teacher. And we achieved a performance of 84.6% on IN1K based on ImageMAE. In summary, our main contribution is as follows:

- We proposed an asymmetric masked distillation strategy for MAE pre-training, which allows the teacher to acquire more context information while maintaining the reconstruction difficulty of the student MAE model.
- We presented a serial feature alignment manner for the asymmetric masking strategy to achieve sufficient knowledge distillation for MAE pre-training.
- We have successfully employed asymmetric masked distillation to obtain the small and robust AMD with improved distillation efficiency and improved transfer performance on downstream tasks.

2. Related work

Vision masked modeling. Recovering an original image from a broken image has recently been introduced as an efficient pre-training paradigm. Early work applied a similar approach for image denoising [55, 56] or image inpainting [26]. Along with the success of Transformer [54] in computer vision, recent works [4, 9, 16, 28, 64, 68] have attempted to apply the Vision Transformer (ViT [17]) to the masked autoencoder. With the success of BERT [15] in proposing a generative task as a pre-training target based on the transformer in NLP, BEiT [4] proposes the masked image modeling based on ViT, which treats image patches as words. The reconstruction target can be divided into high-level features and low-level features. In terms of high-level reconstruction, BEiT performs a two-stage process as the reconstruction target is the discrete token which requires a pre-trained tokenizer [47]. Similarly, PeCO [16] has improved the VAE pre-training by encouraging perceptual similarity.

In terms of low-level reconstruction, Maskfeat [64] proposed the one-stage pre-training approach with a reconstruction target of histograms of oriented gradients (HOG). The reconstruction target of MAE [28] is pixels, and this remarkable work proposes an asymmetric autoencoder structure that leverages a high masking rate to reduce the computational overhead and make scalability possible. SimMIM [68] extends MIM into the Swin Transformer [38] at the cost of a heavier encoder, but which utilises a simple projection head to predict pixels.

Recent works [22, 33, 51, 57] have extended MAE from image to video due to its excellent scalability. VideoMAE [51] applies the tube masking strategy to the video data and adopts joint spatio-temporal attention in the Transformer block to extract video features, which performs excellently on the Something-in-Something V2 [23] dataset. Our work follows VideoMAE and proposes an asymmetric distillation scheme for the pre-trained model.

Knowledge distillation and self distillation. Knowledge distillation (KD) is an efficient way to compress models and was first proposed by Hinton *et al*. in [31]. A common distillation technique is to utilise the logit output from the teacher as a medium for transferring knowledge [2, 8, 31, 45, 52, 67, 70]. And a temperature factor was introduced to align the soft labels with Kullback-Leibler divergence loss. However, the logit-based distillation approach can only be applied to fine-tuned models, which would damage the generalization ability of unsupervised pre-train models.

In addition to logit, researchers have also exploited intermediate features of the model for feature distillation [3, 30, 32, 37, 44, 48, 65, 69, 70]. ViTKD [70] used logit alignment in conjunction with feature alignment and performed well on the fine-tuned model using supervised information. Wei *et al*. [65] have successfully applied feature alignment to the contrastive-based self-supervised learning methods. By analysing the optimization friendliness properties, they conclude that MAE can hardly benefit through the direct feature distillation. dBOT [37] proposed a multi-stage distillation method and use a randomly initialized model as the teacher model. MaskDistill [44] reconstructed the normalized semantic features of the teacher. DMAE [3] adopted a symmetric masking approach based on MAE for feature alignment during student pre-training with pixel reconstruction task. G2SD [34] aligns the features of the decoder and applies two-stage distillation to best exploit the teacher's knowledge. MVD [61] simultaneously employs both image model and video model to distill the MAE model.

A similar style of distillation has emerged in the recent MAE self-distillation works [11, 13, 18, 71]. They constructed the student and teacher models in a two-stream structure, typically designed with the student and teacher masked patch in a complementary relationship. SdAE [13] has successfully accelerated the self-distillation procedure of MAE by analysing information bottleneck and applying the multi-fold masking strategy for the teacher branch. Inspired by those works, our work applies an asymmetric mask strategy to perform feature-based distillation during the pre-training phase of MAE.

3. Method

In this section, we first review the structure of VideoMAE [51], then introduce our AMD framework and introduce the asymmetric mask approach, and finally describe how we perform feature alignment to accomplish the sufficient knowledge distillation.

3.1. Revisiting VideoMAE

VideoMAE [51] extends the masked autoencoder from the image domain to the video domain. Formally, each input video will be randomly sampled into a clip with T frames $V \in \mathbb{R}^{T\times H\times W\times 3}$. The sampling stride τ is set up specifically for the dataset.

Patch embedding. Due to the extra time dimension of the video data, VideoMAE treats a $2 \times 16 \times 16 \times 3$ cube as a patch namely joint space-time cube embedding [1]. Then the 3D-CNN is employed to process the patches, performing convolution without overlap, to obtain a total of $\hat{T} \times \hat{H} \times \hat{W}$ tokens, whose dimension is mapped to D, where $\hat{T} = \frac{T}{2}, \hat{H} = \frac{H}{16}, \hat{W} = \frac{W}{16}$. This allows tokens to be handled in a sequential perspective, whose length is N.

Masking strategy. Due to the redundancy of information in the video data, a higher masking ratio r (*e.g.* 90%) is applied by VideoMAE. To further reduce information leakage, VideoMAE employs tube masking to mask multiple frames. Specifically, a random binary mask map $\tilde{M} \in \mathbb{R}^{\hat{H}\times\hat{W}}$ is first generated in token units. To ensure that a given token in the spatial dimension is masked in all temporal dimensions, VideoMAE simply repeats $\tilde{M}$ in the temporal dimension $\hat{T}$ times to obtain the final mask map $M \in \mathbb{R}^{\hat{T}\times\hat{H}\times\hat{W}}$. We then flatten M into a binary one-dimensional sequence $\hat{M} \in \mathbb{R}^{N\times 1}$ where 1 means that the token needs to be masked and 0 means that it is visible and let P^{vis} denotes the unmasked token indexes.

Encoder: feature extractor. The encoder is a vanilla ViT that takes the sequence of visible tokens after adding the fixed 1D position encoding [54]. It is notable that VideoMAE makes no use of the `[CLS]` token [15]. To allow any two tokens in the entire input sequence to interact with each other, VideoMAE applies the joint space-time attention mechanism [40]. The latent features extracted by the encoder are denoted as $F = \{f^i \in \mathbb{R}^D\}_{i=1}^{\hat{N}}$. When applied to downstream tasks, the encoder acts as a feature extractor.

Decoder: Pixel reconstructor. The task of the decoder is to reconstruct the input, which requires the masked patches to be restored in the form of pixels. The depth of the decoder is usually shallower and less wide than the encoder. A linear layer is used to map the dimension of the latent features F to the width of the decoder D^{dec}. The latent features are then concatenated with the learnable `[MASK]` tokens under the guidance of position and the fixed 1D position encoding is added to them. The decoder is also a vanilla ViT with joint space-time attention, and the output would go through a projection layer to align the dimensions to the original video, which is formed as $\hat{V} \in \mathbb{R}^{T\times H\times W\times 3}$.

Objective function. Following MAE, the pre-training task of VideoMAE is to reconstruct pixels. The loss function applied to the reconstruction is mean square error (MSE) loss, and the reconstruction target is normalised in the token level. The objective function is denoted as:

$$L_{recon} = \frac{1}{rN} \sum_{p \in \bar{P}^{vis}} \left| \mathrm{norm}(V(p)) - \hat{V}(p) \right|^2 , \tag{1}$$

where $\bar{P}^{vis}$ denotes the masked token indexes.

3.2. Overview of AMD

Building on the VideoMAE, we propose an asymmetric masked distillation for the MAE pre-training.

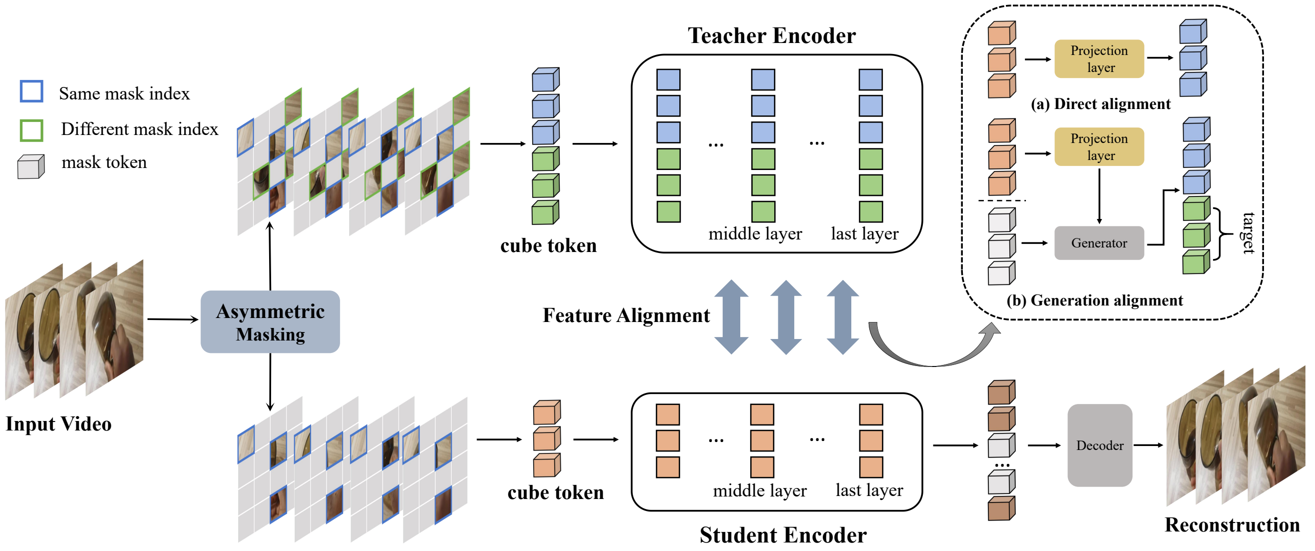


Figure 2. **Pipeline of Asymmetric Masked Distillation (AMD)**. We present an asymmetric masking strategy to transfer the knowledge of teacher pre-trained models to the student masked pre-training. Our asymmetric masking strategy allows a lower masking ratio for the teacher to enable extracting richer visual information. The richer visual information could be used as guidance information to regularize the student masked pre-training and results in a more powerful pre-trained model, that could benefit a variety of downstream tasks.

Overview. The architecture overview is shown in Figure 2. The overall framework of AMD is a two-stream distillation structure with a student branch and a teacher branch, and the teacher is a larger MAE pre-trained model. It is worth noting that AMD only works with the encoder of the teacher for distillation, whereas the students were required to accomplish the pixel reconstruction task with a decoder. Feature distillation occurs between the corresponding layers of the student and teacher models, and AMD employs both direct alignment and generation alignment in a serial way to respond to our asymmetric masking strategy.

Asymmetric mask. A downsampled video clip $V \in \mathbb{R}^{T\times H\times W\times 3}$ is the input of our AMD. After performing the cubic patch embedding, our asymmetric mask strategy is then applied to V, resulting in the input sequence of the student and the teacher, respectively. The length of the input sequence of the student is shorter than that of the teacher, and the visible token index of the student P^{vis}_{stu} is a subset of that of the teacher P^{vis}_{tea}. The details are described in Sec. 3.3.

Serial feature alignment. It is a feature alignment strategy designed for the asymmetric mask, which combines direct alignment and generation alignment with a shared projection function. The two types of alignment are described in Sec. 3.4. Specifically, for a particular layer l of the student, the corresponding layer of the teacher is l^*. We use z to represent the features extracted from ViT the feature can be extracted as z^l_{stu}, $z^{l^*}_{tea}$ respectively.

The first step of direct alignment is to apply the projection function $\phi(\cdot)$ to the student to align the dimension of the teacher. In practice, we employ a linear layer for this purpose. Thus we obtained the student features $\tilde{z}^l_{stu} = \phi\left(z^l_{stu}\right)$ used for direct alignment. We then serially use these features to continue with the generation feature alignment $\mathcal{G}(\tilde{z}^l_{stu})$. This serial alignment can appropriately reduce the difficulty of the generation alignment task.

3.3. Asymmetric Masked Architecture

The asymmetry lies in the difference between the inputs of the student and the teacher. Specifically, the masking ratio of the teacher r_{tea} is lower than the masking ratio of the student r_{stu}. Firstly, we generate the token-wise mask map for the student and teacher, and then get the visible token indexes of the student $P^{vis}_{stu} = \{p^i_{stu}\}^{\hat{N}_{stu}}_{i=1}$ where $\hat{N}_{stu} = (1 - r_{stu})N$ and the teacher $P^{vis}_{tea} = \{p^i_{tea}\}^{\hat{N}_{tea}}_{i=1}$ where $\hat{N}_{tea} = (1 - r_{tea})N$. The relationship between the two of them is formalized as:

$$P^{vis}_{stu} \subsetneqq P^{vis}_{tea}. \tag{2}$$

Thus the teacher can acquire more context information while preserving exactly what the student can receive.

3.4. Feature Distillation

Our work focuses on the pre-training of MAE and employs a serial feature alignment to handle asymmetric masked distillation. Specifically, direct feature alignment and generated feature alignment are performed sequentially. Each of the two types of feature alignment manner is described below.

Direct alignment. Assume that under the condition of symmetric masking strategy, which means that the input lengths $\hat{N}$ of the teacher and student are the same and they share the same set of visible token indexes P^{vis}. The features of the student and the teacher obtained from a pair of corresponding layers are respectively denoted as $z_{stu}^{l} \in \mathbb{R}^{\hat{N} \times D_{stu}}$, $z_{tea}^{l^*} \in \mathbb{R}^{\hat{N} \times D_{tea}}$, where D means the dimension of features, l indicates the layer index of the student model and l^* denotes the layer index corresponding to the student layer (*e.g.* the middle or last layer). The feature dimension of the teacher is generally larger, so a projection function ϕ_d for alignment is necessary. The loss function for a single-layer direct alignment can be defined as:

$$L_{dir} = \frac{1}{\hat{N}} \sum_{p_i \in P^{vis}} \left| z_{tea}^{l^*}(p_i) - \phi_d\left(z_{stu}^{l}(p_i)\right) \right|^2, \tag{3}$$

where $z(p)$ indicates that the feature is extracted from the p-th token of the input sequence.

Generation alignment. The generation alignment can be applied when the input length of the student and teacher do not match. The difference in length between the two is denoted as $\hat{N}_{diff} = \hat{N}_{tea} - \hat{N}_{stu}$. And in terms of the visible token index, we denote $P^{diff} = P_{tea}^{vis} \setminus P_{stu}^{vis}$. We denote the features of the student and the teacher obtained from a pair of corresponding layers by z_{stu}^{l}, $z_{tea}^{l^*}$ respectively. As the feature dimensions of the teacher and the student are not aligned, a simple linear layer $\phi_g(\cdot)$ was employed to map the dimension of the student features.

Since the teacher has more tokens than the student, our work utilises the student model to generate tokens to align with the teacher. Specifically, our generator $\mathcal{G}(\cdot)$ is a decoder-like structure with multi-head self-attention (MHA). Similar to the decoder, we need to concatenate the input with the `[MASK]` token z_m and add the fixed 1D position encoding (PE). The generation process can be described as:

$$\tilde{z}_{stu}^{l} = \phi_g\left(z_{stu}^{l}\right) \in \mathbb{R}^{\hat{N}_{stu} \times D_{tea}}, \tag{4}$$

$$\mathcal{G}\left(\tilde{z}_{stu}^{l}\right) = \text{MHA}\left(\text{concat}\left(\tilde{z}_{stu}^{l}, \text{repeat}\left(z_m\right)\right) + \text{PE}\right). \tag{5}$$

It is worth noting that in a normal decoder structure, the repeat number of the `[MASK]` token is the number of all invisible tokens. However, in our work, we only repeat the `[MASK]` token $\hat{N}_{diff}$ times in order to reduce the generation of redundant features during alignment. We employ the MSE as the loss function for the training of feature alignment for a single layer:

$$L_{gen} = \frac{1}{\hat{N}_{diff}} \sum_{p_i \in P^{diff}} \left| z_{tea}^{l^*}(p_i) - \mathcal{G}\left(\tilde{z}_{stu}^{l}\right)(p_i) \right|^2, \tag{6}$$

3.5. Objective function

Multi-layer alignment. We can choose more than one layer for feature alignment, in practice, we have experimented with two layers, the middle layer and the last layer. As the distribution of features within different layers is varied, the parameters used for alignment are not shared between the different layers. Therefore, when aligning multiple layers, the loss function for direct alignment is rewritten as:

$$L_{dir} = \sum_{l} \frac{1}{\hat{N}_{stu}} \sum_{p_i \in P_{stu}^{vis}} \left| z_{tea}^{l^*}(p_i) - \phi_l\left(z_{stu}^{l}(p_i)\right) \right|^2, \tag{7}$$

where ϕ_l represents the projection function for layer l. The loss function for generation alignment is rewritten as:

$$L_{gen} = \sum_{l} \frac{1}{\hat{N}_{diff}} \sum_{p_i \in P^{diff}} \left| z_{tea}^{l^*}(p_i) - \mathcal{G}_l\left(\tilde{z}_{stu}^{l}\right)(p_i) \right|^2, \tag{8}$$

where $\mathcal{G}_l$ represents the generator for layer l.

Overall. Students are required to reconstruct pixels while completing serial feature alignment, so the overall loss function for AMD is defined as:

$$L_{total} = L_{recon} + L_{dir} + L_{gen}. \tag{9}$$

4. Experiments

4.1. Datasets

Following VideoMAE, we evaluate our AMD on five video datasets: Something-Something V2 (SSV2) [23], Kinetics-400 (K400) [7], UCF101 [49], HMDB51 [35] and AVA [25]. SSV2 contains around 169k training videos and 20k validation videos belonging to 174 action classes. K400 contains about 240k training videos and 20k validation videos from 400 categories. UCF101 and HMDB51 are two small datasets which contain around 9.5k/3.5k training videos and 3.5k/1.5k validation videos respectively. AVA contains 211k training videos and 57k validation videos which is a benchmark for the spatio-temporal localization task. AMD is only pre-trained on SSV2 and K400 and other datasets are used for fine-tuning only. The implementation details are presented in the appendix.

4.2. Ablation Study

In this section, we perform ablation experiments on AMD with 16-frames vanilla ViT-B, and all results are obtained on SSV2. We run 200 epochs per experiment. The ViT-L model of VideoMAE with 2400 epochs pre-trained on SSV2 is employed as the teacher model for distillation.

Since our distillation strategy is built into the MAE pre-training process, we fixed the masking ratio of the student at 90% which is the default setting of VideoMAE, in order not to damage the reconstruction difficulty. As for the fine-tuning setting, the sampling strategy adopted for SSV2 is

Block	Top-1	Time
1	72.06	12.92h
2	72.38	13.33h
4	**72.45**	15.33h

(a) **Generator depth**. Our default choice is a reasonable compromise between performance and computational overhead.

Layers	Stu.	Tea.	Top-1
1	12	24	71.93
2	6	12	**72.38**
	12	24	

(b) **Layer for alignment**. We compare different layer numbers for alignment. Our choice of alignment layers is the middle and the last layer.

Tea. ratio	Top-1	Time
85%	72.17	12.08h
75%	72.38	13.33h
60%	72.44	15.47h
45%	**72.48**	18.93h

(c) **Masking ratio**. Student's ratio is fixed at 90%. Our default choice trade-offs performance and computational overhead.

Method	Top-1	Time
Direct only	71.97	10.83h
Generation only	71.94	12.92h
D+G (in parallel way)	72.25	13.33h
D+G (in serial way)	**72.38**	13.33h

(d) **Distill manner.** AMD works best when using a serial approach combining direct alignment and generative alignment for feature alignment.

Table 1. **Ablation experiments on Something-Something V2.** All models are trained and timed for 200 epochs on 16 A100 GPUs. The student model is the vanilla ViT-B and the teacher model is the ViT-L of videoMAE[51] with 2400 epochs pre-trained on SSV2. The default choice for our AMD is colored in gray.

uniform sampling [59] and a 2 clips ×3 crops test is used to obtain the final results.

Generator depth. The generator is a decoder-like structure and we compared the effect of different generator depths on the performance of the model in Table 1a. A deeper generator provides better classification performance but also brings more computational overhead. The performance gain from 1 layer to 2 layers is significant, but the gain from 2 layers to 4 layers is lower and more time-consuming. We have chosen a depth of 2 as the default setting, which compromises performance and computational complexity.

Alignment layer selection. When performing feature alignment, we only consider the middle and last layers, which for ViT-B are the 6-th and 12-th layer. We have experimented with different alignment layers in Table 1b. When training 200 epochs, we found that aligning two layers simultaneously could produce higher distillation benefits than aligning only one layer. The two are similar in terms of time spent.

Masking ratio. In knowledge distillation, we expect the teacher to transfer more context information to the student, which can be achieved by adjusting the masking ratio of the teacher. The comparison is shown in Table 1c. A lower masking ratio can bring better performance. However, the teacher is the computational bottleneck in the whole structure, and the inference time of the teacher model increases significantly as the masking ratio decreases. So it is unreasonable to reduce the masking ratio hardly. By default, we choose 75% as the masking ratio of the teacher model.

Distill manner. As feature alignment can be divided into direct alignment and generation alignment, they can be used individually or in combination way. We compared the parallel combination strategy with the serial combination strategy in Table 1d, the difference between the two ways is whether the mapping functions are shared or not. We consider that the serial alignment outperforms others as it makes it less difficult for the generation alignment.

Loss function. For the loss function applied for feature alignment, we compared L1 loss and MSE loss. The performance of the former was 71.93% and that of the latter was 72.38%. Therefore, the MSE loss is our default setting.

Method	200 epochs	400 epochs	800 epochs
VideoMAE-B [51]	66.4	67.9	69.6
AMD-B (ours)	**72.4**	**72.8**	**73.3**

Table 2. **The effect of training schedule on SSV2.**

Method	K400 → SSV2	K400 → UCF	K400 → HMDB
VideoMAE-B [51]	68.5	96.1	73.3
AMD-B (ours)	**72.6** (↑ 4.1)	**97.1** (↑ 1.0)	**79.6** (↑ 6.3)

Table 3. **Comparison of the transfer performance.**

4.3. Main Results and Analysis

AMD: small and strong MAE. The aim of our asymmetric distillation is to obtain a smaller yet stronger pre-trained MAE model by feature distillation. As MAE benefits from more training epochs, we experimented AMD at different training epochs in Table 2. Compared with the VideoMAE base model, AMD performs better at a larger interval on different training schedules. It is significant to note that the performance of the teacher model on SSV2 is 74.3%, and our best result at the training schedule of 800 epochs is only 1% lower than it. Therefore, our AMD is a small and strong model thanks to the sufficient distillation.

Transfer learning: action recognition. To verify the generalization ability of the distilled pre-trained model, we pretrained AMD for 800 epochs on K400 and fine-tuned on SSV2, UCF101 and HMDB51. The result is presented in Table 3. Our asymmetric distillation approach significantly improves the generalisation ability of VideoMAE. It is notable that on the SSV2 dataset, the transfer performance is improved by 4.1% compared to VideoMAE, which indicates that AMD reduces the risk of overfitting in transfer fine-tuning and demonstrates robust transfer performance.

Transfer learning: action detection. Following the evaluation settings of VideoMAE, We also transfer the AMD pre-trained with the K400 to the action detection task on AVA in Table 6. With unlabeled data, our AMD could achieve 29.9 mAP. If the intermediate fine-tuning is applied, our AMD could achieve 33.5 mAP. The results show that asymmetric distillation of VideoMAE can also further improve the downstream performance of the non-classification task, which further supports the robustness of our AMD.

	Method	Backbone	Extra data	Extra labels	Frames	GFLOPs	Param	Top-1	Top-5
20-55M # Param	VideoMAE$_{2400e}$ [51]	ViT-S	–	✗	16	57×2×3	22M	66.8	90.3
	MViTv1 [19]	MViTv1-B	Kinetics-400	✓	64	455×1×3	37M	67.7	90.9
	TEINet$_{En}$ [39]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50M	66.5	N/A
	TANet$_{En}$ [42]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×2×3	51M	66.0	90.1
	SlowFast [20]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53M	63.1	87.6
	AMD$_{800e}$ (ours)	ViT-S	–	✗	16	57×2×3	22M	**70.2**	**92.5**
55-100M # Param	VideoMAE$_{800e}$ [51]	ViT-B	–	✗	16	180×2×3	87M	69.6	92.0
	VideoMAE$_{2400e}$ [51]	ViT-B	–	✗	16	180×2×3	87M	70.8	92.4
	Video Swin [41]	Swin-B	IN-21K+K400	✓	32	321×1×3	88M	69.6	92.7
	TDN$_{En}$ [58]	ResNet101$_{\times 2}$	ImageNet-1K	✓	8+16	198×1×3	88M	69.6	92.2
	BEVT [60]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88M	70.6	N/A
	DMAE$^{\dagger}_{800e}$ [3]	ViT-B	–	✗	16	180×2×3	87M	70.0	92.5
	AMD$_{800e}$ (ours)	ViT-B	–	✗	16	180×2×3	87M	**73.3**	**94.0**
>100M # Param	Motionformer [43]	ViT-B	IN-21K+K400	✓	16	370×1×3	109M	66.5	90.1
	TimeSformer [5]	ViT-B	ImageNet-21K	✓	8	196×1×3	121M	59.5	N/A
	ViViT FE [1]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
	VIMPAC [50]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307M	68.1	N/A
	Motionformer [43]	ViT-L	IN-21K+K400	✓	32	1185×1×3	382M	68.1	91.2
	TimeSformer [5]	ViT-L	ImageNet-21K	✓	64	5549×1×3	430M	62.4	N/A
	VideoMAE$_{2400e}$ [51]	ViT-L	–	✗	16	597×2×3	305M	74.3	94.6

Table 4. **Comparison with previous works on SSV2.** Our AMD is pre-trained for 800 epochs on SSV2 using the serial feature alignment strategy. ✗ indicates no additional label information is used for pre-training. ”N/A” means it is not available. ”†” denotes our implementation. DMAE is a distillation method. The teacher model is colored in gray .

	Method	Backbone	Extra data	Extra labels	Frames	GFLOPs	Param	Top-1	Top-5
20-55M # Param	VideoMAE [51]	ViT-S	–	✗	16	57×5×3	22M	79.0	93.8
	TSM [36]	ResNet50	ImageNet-1K	✓	128	65×10×3	24M	74.7	91.4
	MViTv1 [19]	MViTv1-S	–	✗	16	33×5×1	26M	76.0	92.1
	ip-CSN [53]	ResNet152	–	✗	32	109×10×3	33M	77.8	92.8
	NL I3D [63]	ResNet50	ImageNet-1K	✓	128	282×10×3	35M	72.5	90.2
	MViTv1 [19]	MViTv1-B	–	✗	16	71×5×1	37M	78.4	93.5
	AMD$_{800e}$ (ours)	ViT-S	–	✗	16	57×5×3	22M	**80.1**	**94.5**
55-100M # Param	TANet [42]	ResNet152	ImageNet-1K	✓	16	242×4×3	59M	79.3	94.1
	SlowFast [20]	R101+NL	–	✗	16+64	234×10×3	60M	79.8	93.9
	MAE-ST [22]	ViT-B	–	✗	16	180×7×3	87M	81.3	94.9
	VideoMAE$_{800e}$ [51]	ViT-B	–	✗	16	180×5×3	87M	80.0	94.4
	VideoMAE$_{1600e}$ [51]	ViT-B	–	✗	16	180×5×3	87M	81.5	95.1
	TDN$_{En}$ [58]	ResNet101	ImageNet-1K	✓	8+16	198×10×3	88M	79.4	94.4
	BEVT [60]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88M	80.6	N/A
	Video Swin [41]	Swin-B	ImageNet-21K	✓	32	282×4×3	88M	80.6	94.6
	DMAE$^{\dagger}_{800e}$ [3]	ViT-B	–	✗	16	180×5×3	87M	80.8	94.6
	AMD$_{800e}$ (ours)	ViT-B	–	✗	16	180×5×3	87M	**82.2**	**95.3**
>100M # Param	Motionformer [43]	ViT-B	ImageNet-21K	✓	32	370×10×3	109M	79.7	94.2
	TimeSformer [5]	ViT-B	ImageNet-21K	✓	96	590×1×3	121M	78.0	93.7
	ViViT FE [1]	ViT-L	ImageNet-21K	✓	128	3980×1×3	N/A	81.7	93.8
	VIMPAC [50]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307M	77.4	N/A
	Motionformer [43]	ViT-L	ImageNet-21K	✓	32	1185×10×3	382M	80.2	94.8
	TimeSformer [5]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430M	80.7	94.7
	VideoMAE$_{1600e}$ [51]	ViT-L	–	✗	16	597×5×3	305M	85.2	96.8

Table 5. **Comparison with previous works on K400**. Our AMD is pre-trained for 800 epochs on K400 using the serial feature alignment strategy. ✗ indicates no additional label information is used for pre-training. ”N/A” means it is not available. ”†” denotes our implementation. DMAE is a distillation method. The teacher model is colored in gray .

Extreme case: the teacher model performs no masking. When we set the masking ratio of the teacher model to 0%, we tried two settings, one where the student model does not perform any masking either, and the other where the student model remains at 90% masking ratio. We perform the comparison with 200 epochs of training in Table 7.

In the former case, we attempted to align the features directly without any masking of the teacher and the student. As the student is not masked, there is a retreat to the typical feature distillation. However, when the teacher's masking ratio was 45%, AMD took significantly less time to achieve better performance (72.5% vs 72.4%), which illustrates the efficiency of AMD. Since the teacher is the computational bottleneck of distillation, this direct feature alignment without mask imposes a higher computational overhead.

In the latter case, we attempted to adjust the masking ratio for the teacher to 0% in AMD's default settings, which resulted in a relatively poor performance. We believe that

Method	Backbone	Extra labels	$T \times \tau$	mAP
supervised [20]	SlowFast-R101	✓	8×8	23.8
CVRL [46]	SlowOnly-R50	✗	32×2	16.3
ρBYOL$_{\rho=3}$ [21]	SlowOnly-R50	✗	8×8	23.4
ρMoCo$_{\rho=3}$ [21]	SlowOnly-R50	✗	8×8	20.3
VideoMAE [51]	ViT-B	✗	16×4	26.7
VideoMAE [51]	ViT-B	✓	16×4	31.8
AMD (ours)	ViT-B	✗	16×4	**29.9**
AMD (ours)	ViT-B	✓	16×4	**33.5**

Table 6. **Comparison with previous works on AVA v2.2.** "✓" means we perform intermediate fine-tuning on K400 with *labels* before transferred to AVA. $T \times \tau$ denotes the frame number and the sampling rate.

Method	Backbone	Tea.	Stu.	Top-1	Time
No-Masking	ViT-B	0%	0%	72.4	32h
AMD	ViT-B	45%	90%	**72.5**	**19h**
AMD	ViT-B	0%	90%	72.1	26h

Table 7. **Comparison with the teacher without masking.**

the extreme setting would make the generation alignment difficult. We suggest choosing the teacher's masking ratio from efficiency considerations as analyzed in Sec. 4.2. And more analyses can be found in the appendix.

4.4. Application of AMD on Image Model

We apply the asymmetric distillation to the image model ImageMAE [27] on ImageNet-1K [14] to verify the generalizability of AMD. The masking ratio of the student and the teacher is 75% and 50% respectively. The official ViT-B and ViT-L pre-trained models are adopted for the teacher model. The results are presented in Table 8. when we adapt MAE-B as the teacher, AMD can achieve a comparable performance of 82.1% based on ViT-S. when we adapt MAE-L as the teacher, AMD can achieve a classification accuracy of 84.6% based on ViT-B, surpassing ImageMAE by 1.0%.

4.5. Comparison with the Symmetric Method

To reveal the effect of context information in distillation, we compared our AMD with the symmetric method DMAE [3]. In the image domain, in Table 8, AMD outperforms DMAE by 2.8% with MAE-B as the teacher model. And it remains better than 0.6% with MAE-L as the teacher model. In the video domain, in Table 4, with the same teacher model and training length, AMD outperformed DMAE by a margin of 3.3% in the SSV2 dataset. in Table 5, AMD still outperforms DMAE by 1.4% in the K400 dataset. Due to the asymmetric masking, the teacher model could capture more context information, which is beneficial for distillation and the serial alignment exploits context information better. However, the symmetric masking structure merely exploits the stronger feature representation capability of the teacher model. More analyses can be found in the appendix.

Method	Student	Teacher	Top-1 Acc
ImageMAE [27]	ViT-B	-	83.6
ImageMAE [27]	ViT-L	-	85.9
SSTA [66]	DeiT-S	DeiT-B	81.4
DMAE [3]	ViT-S	MAE-B	79.3
G2SD w/o S.D [34]	ViT-S	MAE-B	82.0
AMD (ours)	ViT-S	MAE-B	**82.1**
DMAE [3]	ViT-B	MAE-L	84.0
AMD (ours)	ViT-B	MAE-L	**84.6**

Table 8. **Performance of AMD applied to image model.**

4.6. Comparison with the State of the Art

We compare the previous state-of-the-art results with our AMD on K400 and SSV2 in Table 4 and Table 5 respectively. We pre-trained AMD for 800 epochs for comparison based on both 16-frame vanilla ViT-S and ViT-B. We divided the models into three groups by using 55M and 100M as the boundaries for the number of model parameters. The results show that in the first two groups, AMD achieves the best results with a relatively small number of parameters.

In the group below 55M, our AMD achieves the top-1 accuracy of 70.2% on SSV2 and 80.1% on K400 with no extra data used. Based on ViT-S, our AMD with 800 epochs of pre-training outperforms the VideoMAE with 2400 epochs of pre-training by 3.4% on SSV2 and by 1.1% on K400.

In the group above 55M, our AMD achieves the top-1 accuracy of 73.3% on SSV2 and 82.2% on K400 without extra data used for pre-training. Based on ViT-B, our AMD with 800 epochs of pre-training outperforms the VideoMAE with 800 epochs of pre-training by 3.7% on SSV2 and by 2.2% on K400. It is worth noting that our AMD surpasses some ViT-L based methods [1, 5, 43, 50].

5. Conclusion

In this paper, we have proposed an asymmetric masked distillation framework, termed as AMD, for pre-training relatively small foundation models with autoencoding. The asymmetric masking strategy allows the teacher model to capture more context information to transfer to the student model. We then proposed a customized feature alignment distillation method to take the advantage of the asymmetry, which can better exploit context information. Our AMD can yield smaller foundation models with excellent generalisation capabilities. We apply AMD to both ImageMAE and VideoMAE to demonstrate its effectiveness and versatility, obtaining impressive results in image classification and action recognition with a ViT-B backbone.

Acknowledgements. This work is supported by the National Key R&D Program of China (No. 2022ZD0160900), the National Natural Science Foundation of China (No. 62076119, No. 61921006), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] 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*, pages 6836–6846, 2021. 3, 7, 8

[2] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NIPS*, 2014. 3

[3] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *CVPR*, 2023. 2, 3, 7, 8

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 1, 2

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 7, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5

[8] Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, 2017. 3

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[11] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, pages 1–16, 2023. 3

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 1

[13] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 1, 2, 3

[16] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 552–560, 2023. 2

[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 *International Conference on Learning Representations*, 2020. 2

[18] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 3

[19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021. 7

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF International Conference on Computer Vision*, 2019. 7, 8

[21] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[22] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 1, 2, 7

[23] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 2, 5

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1

[25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 5

[26] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2013. 2

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable

vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 8

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 2

[29] 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*, pages 9729–9738, 2020. 1

[30] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1921–1930, 2019. 3

[31] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[32] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and S. Y. Kung. Milan: Masked image pretraining on language assisted representation. *ArXiv*, abs/2208.06049, 2022. 3

[33] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023. 1, 2

[34] Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15996–16005, 2023. 3, 8

[35] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5

[36] Ji Lin, Chuang Gan, and Song Han. TSM: temporal shift module for efficient video understanding. In *IEEE/CVF International Conference on Computer Vision*, 2019. 7

[37] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 3

[38] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[39] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. TEINet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 7

[40] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2022. 3

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

[42] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *IEEE/CVF International Conference on Computer Vision*, 2021. 7

[43] Mandela Patrick, Dylan Campbell, Yuki 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 *Advances in Neural Information Processing Systems*, 2021. 7, 8

[44] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*, 2022. 3

[45] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 961–969, 2019. 3

[46] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[48] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015. 3

[49] 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

[50] 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. 7, 8

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 6, 7, 8

[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv'e J'egou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 3

[53] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *IEEE/CVF International Conference on Computer Vision*, 2019. 7

[54] 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. 2, 3

[55] 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, pages 1096–1103, 2008. 2

[56] 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. 2

[57] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, June 2023. 1, 2

[58] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: Temporal difference networks for efficient action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 7

[59] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2019. 6

[60] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7

[61] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *CVPR*, 2023. 3

[62] Xiao Wang, Haoqi Fan, Yuandong Tian, Daisuke Kihara, and Xinlei Chen. On the importance of asymmetry for siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16570–16579, 2022. 1

[63] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 7

[64] 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*, pages 14668–14678, 2022. 1, 2

[65] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *ArXiv*, abs/2205.14141, 2022. 2, 3

[66] Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised models are good teaching assistants for vision transformers. In *International Conference on Machine Learning*, pages 24031–24042. PMLR, 2022. 8

[67] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, 2022. 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1, 2

[69] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4633–4642, 2022. 3

[70] Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Practical guidelines for vit feature knowledge distillation. *arXiv preprint arXiv:2209.02432*, 2022. 3

[71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image bert pre-training with online tokenizer. *International Conference on Learning Representations*, 2022. 3