TinyMIM: An Empirical Study of Distilling MIM Pre-trained Models

Sucheng Ren Fangyun Wei* Zheng Zhang Han Hu
Microsoft Research Asia

Abstract

Masked image modeling (MIM) performs strongly in pre-training large vision Transformers (ViTs). However, small models that are critical for real-world applications cannot or only marginally benefit from this pre-training approach. In this paper, we explore distillation techniques to transfer the success of large MIM-based pre-trained models to smaller ones. We systematically study different options in the distillation framework, including distilling targets, losses, input, network regularization, sequential distillation, etc, revealing that: 1) Distilling token relations is more effective than CLS token- and feature-based distillation; 2) An intermediate layer of the teacher network as target perform better than that using the last layer when the depth of the student mismatches that of the teacher; 3) Weak regularization is preferred; etc. With these findings, we achieve significant fine-tuning accuracy improvements over the scratch MIM pre-training on ImageNet-1K classification, using all the ViT-Tiny, ViT-Small, and ViT-base models, with +4.2%/+2.4%/+1.4% gains, respectively. Our TinyMIM model of base size achieves 52.2 mIoU in AE20K semantic segmentation, which is +4.1 higher than the MAE baseline. Our TinyMIM model of tiny size achieves 79.6% top-1 accuracy on ImageNet-1K image classification, which sets a new record for small vision models of the same size and computation budget. This strong performance suggests an alternative way for developing small vision Transformer models, that is, by exploring better training methods rather than introducing inductive biases into architectures as in most previous works. Code is available at https://github.com/OliverRensu/TinyMIM.

1. Introduction

Masked image modeling (MIM), which masks a large portion of the image area and trains a network to recover the original signals for the masked area, has proven to be a very effective self-supervised method for pre-training vision Transformers [2, 11, 17, 49]. Thanks to its strong fine-tuning performance, MIM has now been a main-stream pre-training method for vision Transformers, and numerous follow-ups have been carried out in this research line, such as studying how to set decoding architectures [21], reconstruction targets [10, 32, 45, 59], etc., as well as revealing its properties [46, 48, 50].

*Corresponding author: fawe@microsoft.com.

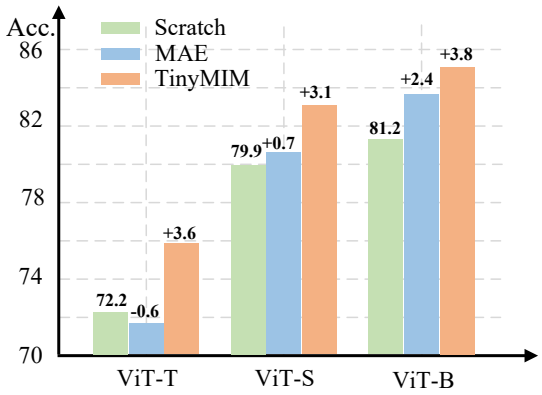

Figure 1. Comparison among TinyMIM (ours), MAE [17] and training from scratch by using ViT-T, -S and -B on ImageNet-1K. We report top-1 accuracy. We adopt DeiT [41] when training from scratch. For the first time, we successfully perform masked image modeling pre-training for smaller ViTs.

Model	Param. (M)	Flops (G)	Top-1 (%)	mIoU
DeiT-T [41]	5.5	1.3	72.2	38.0
PVT-T [43]	13.0	1.9	75.1	39.8
CiT-T [36]	5.5	1.3	75.3	38.5
Swin [29]	8.8	1.2	76.9	40.4
EdgeViT-XS [31]	6.4	1.1	77.5	42.1
MobileViTv1-S [30]	4.9	2.0	78.4	42.7
MobileViTv3-S [42]	4.8	1.8	79.3	43.1
TinyMIM*-T (Ours)	5.8	1.3	**79.6**	**45.0**

Table 1. Comparison with state-of-the-art tiny Transformers with architecture variants. The parameters indicate the backbone parameter excluding the parameters of the last classification layer in classification or the decoder in segmentation. We report top-1 accuracy on ImageNet-1K classification and mIoU on ADE20K segmentation.

Method	ViT-T	ViT-S	ViT-B	ViT-L
Scratch	72.2	79.9	81.2	82.6
MAE	71.6	80.6	83.6	85.9
Gap	-0.6	+0.7	+2.4	+3.3

Table 2. Comparison between MAE pre-trained ViTs and ViTs trained from scratch by using ViT-T, -S, -B and -L on ImageNet-1K. We adopt DeiT when training from scratch. We report top-1 accuracy. As model size shrinks, the superiority of MAE gradually vanishes. MAE even hurts the performance of ViT-T.

However, as shown in Table 2, MIM pre-training [17] mainly effects for relatively large models. When the model size is as small as ViT-Tiny (5 million parameters), which is critical for real-world applications, MIM pre-training can even hurt the fine-tuning accuracy on ImageNet-1K classification. In fact, the accuracy drops by -0.6 compared to the counterpart trained from scratch. This raises a question: can small models also benefit from MIM pre-training, and how can this be achieved?

In addition, the existing study on small vision Transformers mainly focus on introducing certain inductive bias into architecture design [6, 22, 30, 31]. The additional architectural inductive biases facilitate optimization yet limit the expressive capacity. It's natural to ask whether we can boost plain small vision Transformers to perform just as well.

In this work, we present TinyMIM, which answers the above questions. Instead of directly training small ViT models using a MIM pretext task, TinyMIM uses distillation technology [20] to transfer the knowledge of larger MIM pre-trained models to smaller ones. Distillation endows the nice properties of larger MIM pre-trained models to smaller ones while avoiding solving a "too" difficult MIM task. Noting that knowledge distillation has been well developed, especially for supervised models [15], our main work is to systematically study for the first time the effects of different design options in a distillation framework when using MIM pre-trained models as teachers. Specifically, we consider distillation targets, data augmentation, network regularization, auxiliary losses, macro distillation strategy, etc., and draw several useful findings:

- *Distillation targets*. There are two main findings related to distillation targets: 1) Distilling token relations is more effective than distilling the CLS token and feature maps. 2) Using intermediate layers as the target may perform better than using the last layer, and the optimal target layer for different down-stream tasks, e.g., classification and segmentation, can be different.

- *Data and network regularization*. Weak augmentation and regularization is preferred: 1) The performance of using a masked image is worse than using the original image; 2) Relatively small drop path rate (0 for teacher and 0.1 for student) performs best.

- *auxiliary losses*. We find that an auxiliary MIM loss does not improve fine-tuning accuracy.

- *Macro distillation strategy*. We find that using a sequential distillation strategy, i.e., "ViT-B $\rightarrow$ ViT-S $\rightarrow$ ViT-T", performs better than that distilling directly from ViT-B to ViT-T.

By selecting the best framework options, we achieve significant fine-tuning accuracy improvements over the direct MIM pre-training on ImageNet-1K classification, using ViT models of different sizes, as shown in Figure 1. Specifically, the gains of TinyMIM on the ViT-Tiny, ViT-Small, and ViT-base models are +4.2%/+2.4%/+1.4%, respectively.

In particular, our TinyMIM*-T model with knowledge distillation during finetune-tuning achieves a top-1 accuracy of 79.6% on ImageNet-1K classification (see Table 1), which performs better than all previous works that develop small vision Transformer models by introducing architectural inductive biases or smaller feature resolutions. It sets a new accuracy record using similar model size and computation budget. On ADE20K semantic segmentation, TinyMIM*-T achieves 45.0 mIoU, which is +1.9 higher than the second best method, MobileViTv3-S [42]. The strong fine-tuning accuracy by TinyMIM*-T suggests an alternative way for developing small vision Transformer models, that is, by exploring better training methods rather than introducing inductive biases into architectures as most previous works have done.

2. Related Works

2.1. Masked Image Modeling

Inspired by the same idea of masking and reconstruction in BERT [9], BEiT [2] is the pioneer to bring such success to computer vision filed by encoding masked images and predicting masked tokens generated by DALL-E [34]. SimMIM [49] and MAE [17] find that reconstructing RGB pixels results in favorable representations. MAE only encodes the visible tokens and produces reconstructed invisible patches. Recently, a lot of works aim at looking for better supervision. PeCo [10] trains a new tokenizer by enforcing perceptual similarity. iBot [59] and data2vec [1] take exponential moving average (EMA) updated models as tokenizers.

The MIM pre-training performs very well on relatively large models from base size to giant size [28, 49]. However, it will hurt the fine-tuning when the model is as small as the tiny size, probably because MIM task is "too" difficult for a small model. This paper explores how to make small vision Transformer models also benefit from MIM training, through a systematic study of the distillation technology.

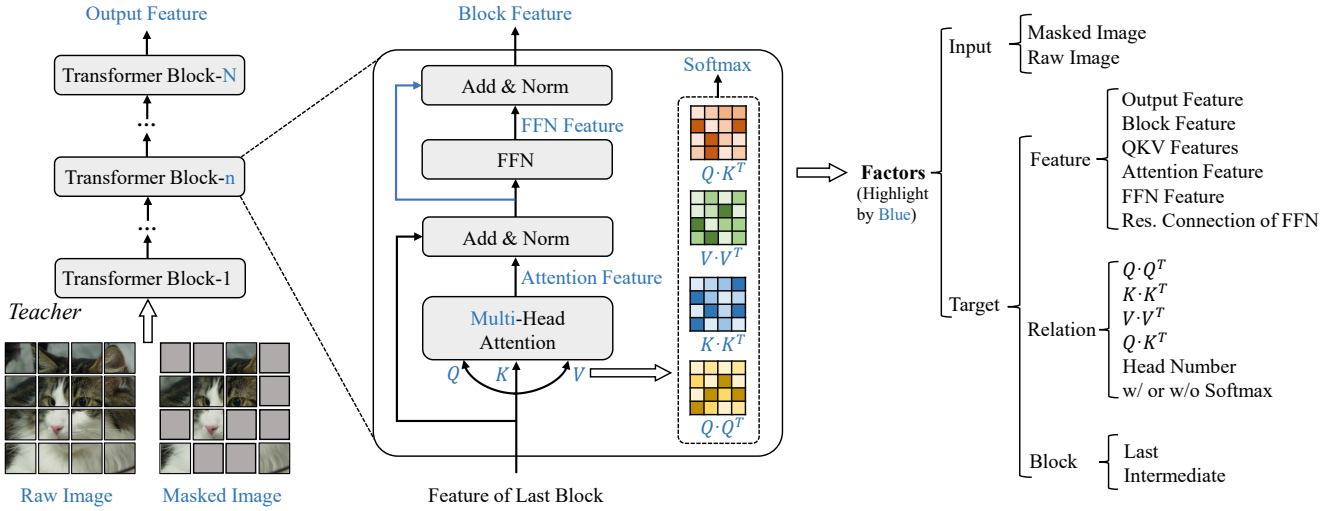

Figure 2. We comprehensively study a variety of factors (highlighted by Royal Blue) that may affect TinyMIM pre-training including input, distillation target (feature or relation) and target block.

2.2. Knowledge Distillation

Knowledge distillation is a classical method to transfer the knowledge from cumbersome models to small ones, pioneered by [20]. The original knowledge distillation framework adopts the annealed classification logits of the teacher as the distilling target for the student. Since then, extensive variants have been carried out to improve the distilling effectiveness [15], including changing the distilling targets as intermediate features [18, 19, 24, 38] and relations [25, 56], data augmentations of teacher and students [36, 47], regularization [47], distilling strategies [27, 35, 44, 51–55, 57, 58] and so on.

While almost all studies are made for CNN architectures under supervised settings, recently, there have been a few works performing distilling technologies for vision Transformers [41, 47] and contrastive learning based methods [13, 47]. In DeiT [41], the teacher is set as a CNN architecture so as to transfer the inductive bias involved in CNNs to vision Transformers. It also proposes to use hard distillation which uses hard pseudo-class labels of the teacher network as the distilling targets, which performs better than the naive knowledge distillation [20]. In [47], a method based on feature map distillation is proposed to generally improve vision transformers by different pre-training approaches including image classification, instance contrastive based self-sueprvised learning [3] and CLIP pre-training [33]. However, it shows no gains for MIM pre-trained models. This paper for the first time studies the distillation framework for MIM pre-trained vision Transformers in which significant gains are achieved for vision Transformers of various sizes.

2.3. Small Vision Transformers

Designing efficient CNN models [23, 39] has been widely studied in recent years. With the emergence of Vision Transformer (ViT) [11], there have been several works studying how to develop efficient vision Transformer, with the majority focus on introducing inductive biases into the architectures [16, 22, 26, 30, 31, 37]. Different from these works that develop small vision Transformers by introducing sophisticated components into architectures, we demonstrate that a plain vision Transformer [11] at a small scale can perform just as well, or even better. Our main insight is that the MIM pre-training can implicitly incorporate necessary inductive biases, and thus avoids the need for explicit architecture bias. Our plain vision Transformer of tiny size achieves state-of-the-art accuracy for both ImageNet-1K image classification and ADE20K semantic segmentation using a similar model size and computation budget.

3. TinyMIM

We adopt a larger, MIM pre-trained model as the teacher, and a smaller ViT as the student. The objective of TinyMIM is to train the randomly initialized student by mimicking the target produced by the teacher in a knowledge distillation manner. After pre-training, the TinyMIM pre-trained model can be transferred to various downstream tasks. In this work, we adopt MAE [17] as the MIM model due to its popularity and simplicity. In this section, we first describe the factors that may affect TinyMIM pre-training: distillation target in Section 3.1.1; input in Section 3.1.2; target block in Section 3.1.3. Then we present a series of distillation losses for different distillation targets in Section 3.1.3. At last, a sequential distillation strategy is introduced to facilitate the

performance in Section 3.3.

3.1. Factors

3.1.1 Distillation Target

Block Feature and Output Feature. Given an input image $\boldsymbol{x}$, we first divide it into N non-overlapping patches and use a linear projection layer to map N patches into patch embeddings $F_0 \in \mathbb{R}^{N\times D}$, where D is the dimension of hidden features. Suppose we have a ViT containing L Transformer blocks. Each Transformer block takes the output F_{i-1} of the last Transformer block as the input and generates the feature F_i of the current block, which can be formulated as:

$$F_i = \text{Transformer}(F_{i-1}), i \in [1, L]. \tag{1}$$

We term F_i as the block feature of the i-th Transformer block. In particular, we name the feature F_L from the last Transformer block as the output feature.

Attention Feature and FFN Feature. Each Transformer block is composed of a self-attention layer and a feed forward layer, which can be defined as:

$$\begin{aligned} H_i &= \text{Attention}(\text{LN}(F_{i-1})), \\ \widehat{H}_i &= H_i + F_{i-1}, \\ \widetilde{H}_i &= \text{FFN}(\text{LN}(\widehat{H}_i)), \\ \overline{F}_i &= \widehat{H}_i + \widetilde{H}_i, \end{aligned} \tag{2}$$

where $\text{Attention}(\cdot)$, $\text{FFN}(\cdot)$ and $\text{LN}(\cdot)$ denotes self-attention layer, feed forward layer and layer norm, respectively. We term $\widehat{H}_i$ and $\widetilde{H}_i$ as attention feature and FFN feature of the i-th Transformer block.

Query/Key/Value Features. Each self-attention layer consists of M head networks, each of which maps input feature F_{i-1} to query (Q), key (K) and value (V):

$$\begin{aligned} Q_i^m &= LN(F_{i-1})W_i^Q, \\ K_i^m &= LN(F_{i-1})W_i^K, \\ V_i^m &= LN(F_{i-1})W_i^V, \end{aligned} \tag{3}$$

where $Q_i, K_i, V_i \in \mathbb{R}^{N\times \frac{D}{M}}$ represent the query, key and value of the m-th head network. The query/key/value features ($Q_i, K_i, V_i \in \mathbb{R}^{N\times D}$) are the concatenation of M Q_i^m/K_i^m/V_i^m, respectively.

Relations. For the m-th head network from the i-th Transformer block, we could calculate its Q-Q, K-K, V-V and Q-K relations ($R_{i,m}^{QQ}, R_{i,m}^{KK}, R_{i,m}^{VV}, R_{i,m}^{QK} \in \mathbb{R}^{N\times N}$), which are implemented as the scaled product relation:

$$\begin{aligned} R_{i,m}^{QQ} &= \text{Softmax}\left(\frac{Q_i^m Q_i^{m\mathsf{T}}}{\sqrt{D/M}}\right), \\ R_{i,m}^{KK} &= \text{Softmax}\left(\frac{K_i^m K_i^{m\mathsf{T}}}{\sqrt{D/M}}\right), \\ R_{i,m}^{VV} &= \text{Softmax}\left(\frac{V_i^m V_i^{m\mathsf{T}}}{\sqrt{D/M}}\right), \\ R_{i,m}^{QK} &= \text{Softmax}\left(\frac{Q_i^m K_i^{m\mathsf{T}}}{\sqrt{D/M}}\right). \end{aligned} \tag{4}$$

The Q-Q/K-K/V-V/Q-K relations (R_i^{QQ}, R_i^{KK}, R_i^{VV}, $R_i^{QK} \in \mathbb{R}^{M\times N\times N}$) of the i-th Transformer block is the stack of M $R_{i,m}^{QQ}$/$R_{i,m}^{KK}$/$R_{i,m}^{VV}$/$R_{i,m}^{QK}$, respectively.

3.1.2 Input

MIM models randomly mask a high proportion of image patches on an input image $\boldsymbol{x}$, yielding a masked image $\widetilde{\boldsymbol{x}}$ for pre-training. We also investigate the input of TinyMIM when performing knowledge distillation— the input could be either a raw image $\boldsymbol{x}$ or a masked image $\widetilde{\boldsymbol{x}}$.

3.1.3 Target Block

Consider a situation where we tend to use an MAE pre-trained ViT-L (teacher) containing 24 blocks to distill a ViT-B (student) containing 12 blocks. In this scenario, the block number of the student does not match that of the teacher. We investigate which block of the teacher can provide the most appropriate target. The selected block is referred to as the target block.

3.2. Knowledge Distillation as MIM Pre-training

In Section 3.1.1, we describe a variety of distillation target candidates. In this section, we introduce different knowledge distillation losses for various distillation targets. Let $\boldsymbol{x}$ denote an input image, f_t and f_s represent a teacher model and a student model, respectively. The objective of knowledge distillation is to transfer the knowledge from f_t to f_s by optimizing f_s while freezing f_t. In general, the training is supervised by the KL divergence, which is defined as:

$$\mathcal{L}_{KL}(\boldsymbol{p}, \boldsymbol{t}) = \boldsymbol{t} log\frac{\boldsymbol{t}}{\boldsymbol{p}}, \tag{5}$$

where $\boldsymbol{t}$ denotes the target generated by $f_t(\boldsymbol{x})$, and $\boldsymbol{p}$ is the prediction produced by $f_s(\boldsymbol{x})$.

Class Token Distillation. We use $\boldsymbol{c}_t$ and $\boldsymbol{c}_s$ to denote class token feature of f_t and f_s, respectively. The loss of class token distillation is formulated as:

$$\mathcal{L} = \mathcal{L}_{KL}(\boldsymbol{c}_s, \boldsymbol{c}_t). \tag{6}$$

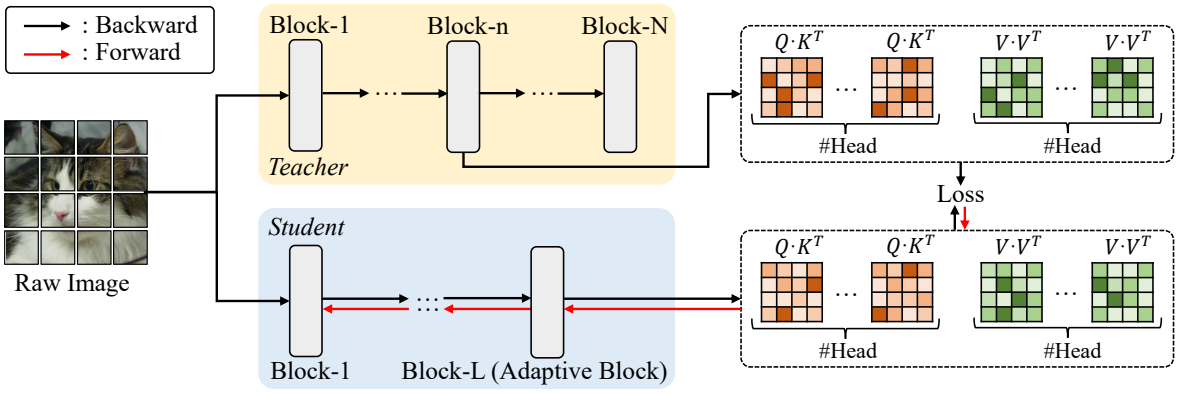

Figure 3. The default knowledge distillation strategy of TinyMIM. The student (*e.g.* ViT-B) is optimized to mimic the relations generated by the intermediate block of a MIM pre-trained teacher (*e.g.* ViT-L) with raw image as input. We replace the last block of the student with an adaptive block to match teacher's head number (no extra computational cost). After pre-training (knowledge distillation), the student model can be transferred to various downstream tasks.

Feature Distillation. In general, the feature dimension of the teacher network and the student network are mismatched. To tackle this problem, we adopt an extra linear layer on the output of the student network to match the feature dimension of the teacher's target. Let $\boldsymbol{F}_t$ and $\boldsymbol{F}_s$ denote the target feature and the prediction yielded by the student followed by a linear projection layer, respectively. We could formulate the loss of feature distillation as follows:

$$\mathcal{L} = \mathcal{L}_1(\boldsymbol{F}_s, \mathrm{Norm}(\boldsymbol{F}_t)), \tag{7}$$

where $\mathrm{Norm}(\cdot)$ is the whitening operation implemented by layer norm without affiliation, and $\mathcal{L}_1$ is the smooth L1 loss defined as:

$$\mathcal{L}_1(y, \hat{y}) = \begin{cases} \frac{1}{2}(\hat{y}-y)^2/\beta, & |\hat{y}-y| \le \beta \\ (|\hat{y}-y| - \frac{1}{2}\beta), & \text{otherwise} \end{cases}, \tag{8}$$

where β is set to 2.0.

Relation Distillation. This is our default knowledge distillation strategy as illustrated in Figure 3. For the sake of clarity, we use $R^{QK}_{t\to m}$ to denote the m-th head generated Q-K relation target (see Eq 4) from the teacher network, and $R^{QK}_{s\to m}$ to represent the corresponding Q-K relation prediction from the student network. We define $R^{VV}_{t\to m}$ and $R^{VV}_{s\to m}$ in a similar way. The loss of relation distillation is formulated as:

$$\begin{aligned} \mathcal{L}^{QK} &= \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{KL}(R^{QK}_{s\to m}, R^{QK}_{t\to m}), \\ \mathcal{L}^{VV} &= \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{KL}(R^{VV}_{s\to m}, R^{VV,S}_{t\to m}), \\ \mathcal{L} &= \mathcal{L}^{QK} + \mathcal{L}^{VV}. \end{aligned} \tag{9}$$

Head Alignment for Relation Distillation. In general, the head number of the student network is lower than that of the teacher network. For instance, ViT-L (teacher) contains 16 heads per block while ViT-B (student) only contains 12 heads per block. Recall that the relation distillation loss (Eq. 9) is calculated head by head, thus we have to solve the head misalignment issue before performing relation distillation. To this end, we replace the last block of the student with an adaptive block, which keeps the original hidden dimension but adjusts the head number to the teacher. Concretely, given a teacher network with M_t heads per block, and a student network with M_s heads per block, a hidden dimension of D_s, and a head dimension of D_s/M_s, the adaptive block is designed to be a Transformer block with M_t heads per block, a hidden dimension of D_s and a head dimension of D_s/M_t.

3.3. Sequential Distillation

When training a small model like ViT-S, the teacher has two options: a pre-trained ViT-B and a pre-trained ViT-L. Intuitively, the pre-trained ViT-L is a good teacher due to its higher representation capability. However, there is a huge capacity gap between ViT-L and ViT-S, resulting in poor distillation results. Following [8, 14], we adopt a sequential distillation strategy to improve pre-training. For instance, when pre-training a ViT-S, the teacher is selected as a TinyMIM pre-trained ViT-B, which has been trained by TinyMIM with ViT-L as the teacher.

4. Experiments

The pretraining and fine-tuning details can be found in supplementary materials.

Default Setting. By default, we adopt relation distillation formulated in Eq. 9, head alignment, raw image as input, sequential distillation and the 18-th block of MAE pre-trained ViT-L as the target block for TinyMIM-ViT-B pre-training.

Method	Pretraining Epochs	Tokenizer/ Teacher	Tokenizer/Teacher Data	Classification Top-1 Acc (%)	Segmentation mIoU
Tiny-size models (ViT-T/16)					
Scratch [41]	300	Label	IN1K	72.2	38.0
MAE† [17]	1600	Pixel	IN1K	71.6	37.6
MoCo [5]	1600	EMA	IN1K	73.3	39.3
TinyMIM (Ours)	300	TinyMIM-ViT-S	IN1K	**75.8**	**44.0/44.6**‡
TinyMIM⋆ (Ours)	300	TinyMIM-ViT-S	IN1K	**79.6**	**45.0**‡
Small-size models (ViT-S/16)					
Scratch [41]	300	Label	IN1K	79.9	43.1
MAE† [17]	1600	Pixel	IN1K	80.6	42.8
MoCo [5]	1600	EMA	IN1K	81.4	43.9
DINO [3]	1600	EMA	IN1K	81.5	45.3
CIM [12]	1600	Pixel	IN1K	81.6	-
TinyMIM (Ours)	300	TinyMIM-ViT-B	IN1K	**83.0**	**48.4/48.9**‡
Base-size models (ViT-B/16)					
Scratch [41]	300	Label	IN1K	81.2	47.2
BeiT [2]	800	DALL-E	DALLE250M+IN22K+IN1K	83.2	45.6
MAE [17]	1600	Pixel	IN1K	83.6	48.1
SIM [40]	1600	EMA	IN1K	83.8	-
CAE [4]	1600	DALL-E	DALLE250M+IN22K+IN1K	83.9	50.2
MaskFeat [45]	1600	HOG	IN1K	84.0	-
SdAE [7]	300	EMA	IN1K	84.1	48.6
data2vec [1]	800	EMA	IN1K	84.2	-
PeCo [10]	300	VQGAN	IN1K	84.1	46.7
PeCo [10]	800	VQGAN	IN1K	84.5	48.5
TinyMIM (Ours)	300	MAE-ViT-L	IN1K	**85.0**	**52.2/52.6**‡

Table 3. Fine-tuning results on ImageNet-1K and ADE20K. All models are pre-trained on ImageNet-1K. "Tokenizer/Teacher Data": training data of teacher and tokenizer. †: reproduced result using official code. ⋆: the model is fine-tuned for 1000 epochs with DeiT-style [41] knowledge distillation. ‡: the model adopts an intermediate fine-tuning on ImageNet-1K classification before ADE20K segmentation fine-tuning.

Method	Model Size	ImageNet ↑	IN-Adversarial↑	IN-Rendition↑	IN-Corruption ↓
DeiT [41]	ViT-T	72.2	8.0	32.7	54.0
MAE [17]		71.8	7.0	36.5	55.2
TinyMIM		**75.8**	**11.0**	**39.8**	**50.1**
DeiT [41]	ViT-S	79.9	18.3	42.3	41.4
MAE [17]		80.6	20.1	45.6	40.6
TinyMIM		**83.0**	**27.5**	**48.8**	**35.8**
DeiT [41]	ViT-B	81.2	25.8	45.4	36.8
MAE [17]		83.6	33.6	50.0	37.8
TinyMIM		**85.0**	**43.0**	**54.6**	**32.7**

Table 4. Robustness evaluation on out-of-domain datasets.

4.1. Main Results

As shown in Table 3, we compare our TinyMIM with previous methods on ImageNet image classification and ADE20K semantic segmentation using different ViTs. In particular, TinyMIM pre-trained ViT-T achieves 75.8% top-1 accuracy, outperforming MAE baseline by **+4.2**. An enhanced model named TinyMIM⋆-T, which retains the plain architecture and computation budget of ViT-T, further achieves 79.6% top-1 accuracy. See appendix for the details of TinyMIM⋆-T. Moreover, TinyMIM pre-trained ViT-S achieves 83.0% top-1 accuracy, outperforming MAE baseline and previous best method CIM [12] by **+2.4**, **+1.4**, respectively. By transferring the knowledge of an MAE pre-

Method	Reconstruction Loss	Top-1 Acc.
MAE	✓	83.6
TinyMIM w/ Cls		80.6
TinyMIM w/ Cls	✓	82.1

Table 5. Study of class token distillation formulated in Eq.6.

Feature	Res. Connection	Top-1 Acc.
MAE		83.6
Output Feature		83.7
FFN Feature		84.2
FFN Feature	✓	81.8
Attention Feature		84.1
Attention Feature	✓	81.3
Q/K/V Features		84.3

Table 6. Study of feature distillation formulated in Eq.7. See Section 3.1.1 and Eq. 2 for the definitions of different features.

trained ViT-L, TinyMIM pre-trained ViT-B achieves 85.0% top-1 accuracy on ImageNet-1K.

As for semantic segmentation, TinyMIM pre-trained ViT-B surpasses MAE baseline and state-of-the-art CAE [4] by **+4.1** and **+2.0**, respectively. An intermediate fine-tuning on ImageNet-1K classification before ADE20K segmentation fine-tuning further boosts the performance.

We also evaluate our models on out-of-domain datasets in Table 4. Our TinyMIM pretrained models are more robust than MAE pre-trained ones. Specifically, TinyMIM-ViT-B outperforms MAE-ViT-B by **+6.4** and **+4.6** on ImageNet-A and ImageNet-R, respectively, and lower the mCE by **-5.1**.

4.2. Ablation Study

Unless otherwise specified, all ablation studies are conducted on TinyMIM-ViT-B, with a teacher of being an MAE pre-trained ViT-L, relation distillation strategy, raw image as input, the 18-th block of ViT-L as the target block, under a 100-epoch pre-training schedule. We report top-1 accuracy on ImageNet-1K.

Class Token Distillation. For this distillation strategy, we study two variants: 1) class token distillation as formulated in Eq.6; 2) class token distillation with an extra MAE reconstruction loss. The results are shown in Table 5. Both variants perform worse than MAE baseline, indicting that the class token is improper to be served as the distillation target since there is no explicit supervision applied on class token during teacher's pre-training.

Feature Distillation. As described in Section 3.1.1, there are four types of features can be served as the targets for feature distillation formulated in Eq. 7: output feature, FFN feature, attention feature and Q/K/V features. Table 6 compares the results of using different features as distillation targets. We also report the results of FFN feature and attention feature before the residual connection (see Eq. 2). An interesting finding is that distilling FFN feature and attention feature after the residual connection significantly degrades the performance.

Relation	Softmax	Top-1 Acc.
MAE		83.6
Q-Q, K-K, V-V		84.4
Q-Q, K-K, V-V	✓	84.5
Q-K, V-V		84.4
Q-K, V-V	✓	84.6

Table 7. Study of relation distillation formulated in Eq. 9. See Section 3.1.1 and Eq. 4 for the definitions of different relations.

Method	Model Size	Top-1 Acc.
Supervised (DeiT)	ViT-T	72.2
MAE		71.6
Class Token Distillation		70.6
Feature Distillation		73.4
Relation Distillation		75.8 (**+4.2**)
Supervised (DeiT)	ViT-S	79.9
MAE		80.6
Class Token Distillation		79.6
Feature Distillation		80.8
Relation Distillation		83.0 (**+3.1**)
Supervised (DeiT)	ViT-B	81.2
MAE		83.6
Class Token Distillation		82.6
Feature Distillation		83.8
Relation Distillation		85.0 (**+1.6**)

Table 8. Comparison of three distillation strategies on ImageNet-1K image classification. The models are pre-trained under a 300-epoch schedule.

Relation Distillation. Eq. 9 formulates our default relation distillation, which jointly distills Q-K relation and V-V relation (see Eq. 4). Here we study a variant by changing the target relations from Q-K/V-V to Q-K/K-K/V-V. We also investigate that whether to apply a Softmax operator on each relation. The results are shown in Table 7.

Comparison of Different Distillation Strategies. In this study, all models are pre-trained under a 300-epoch schedule. We compare three distillation strategies on ImageNet image classification (Table 8) and ADE20K semantic segmentation (Table 9). For each strategy, we use the target that yields the best result. We also highlight the improvements over the MAE baseline.

Target Block. As described in Section 3.1.3, we consider a situation where the block number of the student does not

Method	Model Size	mIoU
Supervised (DeiT)	ViT-B	47.2
MAE		48.1
Class Token Distillation		46.2
Feature Distillation		47.7
Relation Distillation		52.2 (**+4.1**)

Table 9. Comparison of three distillation strategies on ADE20K semantic segmentation. The models are pre-trained under a 300-epoch schedule.

Task	12_{th}	15_{th}	18_{th}	21_{th}	24_{th}
Classification	83.6	84.1	84.6	84.8	84.4
Segmentation	48.7	49.8	52.2	50.6	50.0

Table 10. Study of target block on ImageNet-1K and ADE20K.

Student	Teacher	Acc.
ViT-S	MAE-ViT-B	82.3
	MAE-ViT-L	82.1
	MAE-ViT-L $\rightarrow$ TinyMIM-ViT-B	82.6
ViT-T	MAE-ViT-S	74.1
	MAE-ViT-B	74.4
	MAE-ViT-B $\rightarrow$ TinyMIM-ViT-S	75.0

Table 11. Study of sequential distillation.

match that of the teacher. Here we use an MAE pre-trained ViT-L containing 24 blocks to distill a ViT-B containing 12 blocks. Here we examine the effects of using the 12_{th}, 15_{th}, 18_{th}, 21_{th} and 24_{th} (last) blocks of the ViT-L as the target blocks. The comparison is shown in Table 10. We experimentally find that using 18_{th} block yields the best result.

Sequential Distillation. In Section 3.3, we advocate to adopt a sequential distillation strategy to enable distillation from a larger model (*e.g.* ViT-L) to a smaller model (*e.g.* ViT-S). Table 11 compares the result of adopting different teachers with or without the sequential distillation. We have two conclusions: 1) using a larger teacher (MAE-ViT-L) to distill a smaller student (ViT-S) degrades the performance; 2) sequential distillation significantly boosts the performance of ViT-T (MAE-ViT-B$\rightarrow$TinyMIM-ViT-S as the teacher and ViT-T as the student).

Integrating MAE into TinyMIM. MAE is a simple but effective self-supervised pre-training paradigm that trains a model by requiring it to predict masked inputs. In contrast, TinyMIM pre-trains smaller ViTs in a knowledge distillation manner. Here we integrate MAE into our TinyMIM, yielding an integrated model. This model is optimized under two losses: knowledge distillation loss from TinyMIM, and reconstruction loss from MAE. To enable MAE pre-training,

Masked Image	Reconstruction Loss	Top-1 Acc.
		84.6
✓		83.9
✓	✓	84.0

Table 12. Comparison between the TinyMIM-ViT-B (the first row) and the integrated model (the third row). We also study the input of TinyMIM-ViT-B, which could be raw image (the first row) or masked image (the second row).

DPR (Teacher)	DPR (Student)	Top-1 Acc.
0.0	0.0	84.3
0.0	0.1	84.6
0.0	0.2	84.3
0.0	0.3	84.1
0.1	0.1	83.9

Table 13. Ablation study of drop path rate (DPR) used in teacher and student.

we randomly mask 75% image patches, and feed the visible patches into the network to initiate the pre-training of the integrated model. Table 12 shows the comparison between TinyMIM-ViT-B and the integrated model. From the Table, we could draw a conclusion—integrating MAE into our TinyMIM does not improve the performance. In addition, we also investigate the input of TinyMIM-ViT-B, which could be either raw image or masked image, as shown in Table 12—taking raw image as input yields better result.

Drop Path. Drop path is one of the most critical techniques in training Transformers [41]. Using an appropriate drop path rate could significantly alleviate the over-fitting issue. However, MAE disables this technique in its implementation. Here we verify the effects of applying drop path to our TinyMIM. The results are shown in Table 13. For the student model, the optimal drop path rate is 0.1. For the teacher model, disabling drop path yields best result.

5. Conclusion

In this paper, we present TinyMIM, which is the first to successfully make small models benefit from masked image modeling (MIM) pre-training. Instead of adopting a mask-and-predict pretext task, we pre-train a small ViT by mimicking the relations of a large ViT in a knowledge distillation manner. The success of TinyMIM can be attributed to a comprehensive study of various factors that may affect TinyMIM pretraining including distillation target, distillation input and target block. With extensive experiments, we conclude that relation distillation is superior than feature distillation and class token distillation, etc. With its simplicity and strong performance, we hope our approach can serve as a solid baseline for future research.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 2, 6

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1, 2, 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3, 6

[4] 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. *arXiv preprint arXiv:2202.03026*, 2022. 6, 7

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *ArXiv*, abs/2104.02057, 2021. 6

[6] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022. 2

[7] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. *ArXiv*, abs/2208.00449, 2022. 6

[8] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[10] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 1, 2, 6

[11] 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. *preprint arXiv:2010.11929*, 2020. 1, 3

[12] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 6

[13] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. 2021. 3

[14] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 5

[15] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021. 2, 3

[16] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3, 6

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

[19] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019. 3

[20] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 3

[21] 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. 1

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2, 3

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[24] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3

[25] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–350, 2018. 3

[26] Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.01191*, 2022. 3

[27] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. *arXiv preprint arXiv:2210.16774*, 2022. 3

[28] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[29] 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. *arXiv preprint arXiv:2103.14030*, 2021. 1

[30] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[31] Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *European Conference on Computer Vision*, pages 294–311. Springer, 2022. 1, 2, 3

[32] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[34] A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 2

[35] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13325–13333, June 2021. 3

[36] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16773–16782, June 2022. 1, 3

[37] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10853–10862, June 2022. 3

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 3

[39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[40] Chenxin Tao, Xizhou Zhu, Gao Huang, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022. 6

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020. 1, 3, 6, 8

[42] Shakti N Wadekar and Abhishek Chaurasia. Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*, 2022. 1, 2

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 1

[44] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018. 3

[45] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 1, 6

[46] 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 preprint arXiv:2205.14141*, 2022. 1

[47] 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 preprint arXiv:2205.14141*, 2022. 3

[48] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022. 1

[49] 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

[50] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. *arXiv preprint arXiv:2206.04664*, 2022. 1

[51] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022. 3

[52] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021. 3

[53] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *Computer Vision–ECCV*

2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV, pages 73–91. Springer, 2022. 3

[54] Jingwen Ye, Yixin Ji, Xinchao Wang, Xin Gao, and Mingli Song. Data-free knowledge amalgamation via group-stack dual-gan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12513–12522, 2020. 3

[55] Jingwen Ye, Yining Mao, Jie Song, Xinchao Wang, Cheng Jin, and Mingli Song. Safe distillation box. In *AAAI Conference on Artificial Intelligence*, 2021. 3

[56] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017. 3

[57] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 3

[58] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning with single-teacher multi-student. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[59] 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. 1, 2