Masked Autoencoders Enable Efficient Knowledge Distillers

Yutong Bai[1] Zeyu Wang[2] Junfei Xiao[1] Chen Wei[1]
Huiyu Wang[1] Alan Yuille[1] Yuyin Zhou[2] Cihang Xie[2]
[1]Johns Hopkins University [2] University of California, Santa Cruz
ytongbai@gmail.com

Abstract

This paper studies the potential of distilling knowledge from pre-trained models, especially Masked Autoencoders. Our approach is simple: in addition to optimizing the pixel reconstruction loss on masked inputs, we minimize the distance between the intermediate feature map of the teacher model and that of the student model. This design leads to a computationally efficient knowledge distillation framework, given 1) only a small visible subset of patches is used, and 2) the (cumbersome) teacher model only needs to be partially executed, i.e., forward propagate inputs through the first few layers, for obtaining intermediate feature maps.

Compared to directly distilling fine-tuned models, distilling pre-trained models substantially improves downstream performance. For example, by distilling the knowledge from an MAE pre-trained ViT-L into a ViT-B, our method achieves 84.0% ImageNet top-1 accuracy, outperforming the baseline of directly distilling a fine-tuned ViT-L by 1.2%. More intriguingly, our method can robustly distill knowledge from teacher models even with extremely high masking ratios: e.g., with 95% masking ratio where merely TEN patches are visible during distillation, our ViT-B competitively attains a top-1 ImageNet accuracy of 83.6%; surprisingly, it can still secure 82.4% top-1 ImageNet accuracy by aggressively training with just FOUR visible patches (98% masking ratio). The code and models are publicly available at `https://github.com/UCSC-VLAA/DMAE`.

1. Introduction

Following the success in the natural language processing [10, 27], the Transformer architecture is showing tremendous potentials in computer vision [2, 11, 25, 26, 33], especially when they are pre-trained with a huge amount of unlabelled data [3] with self-supervised learning techniques [1, 14]. Masked image modeling, which trains models to predict the masked signals (either as raw pixels or as semantic tokens) of the input image, stands as one of the most powerful ways for feature pre-training. With the most recent representative work in this direction, masked autoencoder (MAE) [13], we are now able to efficiently and effectively pre-train high-capacity Vision Transformers (ViTs) with strong feature representations, leading to state-of-the-art solutions for a wide range of downstream visual tasks.

In this paper, we are interested in applying knowledge distillation [17], which is one of the most popular model compression techniques, to transfer the knowledge from these strong but cumbersome ViTs into smaller ones. In contrast to prior knowledge distillation works [17, 21, 34], the teacher considered here is a pre-trained model whose predictions do not necessarily reveal the fine-grained relationship between categories; therefore, typical solutions like aligning the soft/hard logits between the teacher model and the student model may no longer remain effective. Moreover, after distilling the pre-trained teacher model, these student models need an extra round of fine-tuning to adapt to downstream tasks. These factors turn distilling pre-trained models seemingly a less favorable design choice in terms of both performance and computational cost.

Nonetheless, surprisingly, we find by building upon MAE, the whole distillation framework can efficiently yield high-performance student models. There are two key designs. Firstly, we follow MAE to let the encoder exclusively operate on a small visible subset of patches and to employ a lightweight decoder for pixel reconstruction. Whereas rather than using the "luxury" setups in MAE, we show aggressively *simplifying pre-training from 1600 epochs to 100 epochs* and *pushing masking ratio from 75% to 95%* suffice to distill strong student models. Secondly, instead of aligning logits, we alternatively seek to match the intermediate feature representation; this enables the cumbersome teacher model to only forward propagate inputs through the first few layers, therefore, reducing computations. We note applying *L1 norm for distance measure* is essential recipe for ensuring a successful intermediate feature alignment.

We name this distilling MAE framework as DMAE. Compared to the traditional knowledge distillation framework where the teacher is a fine-tuned model, DMAE is more efficient and can train much stronger student models

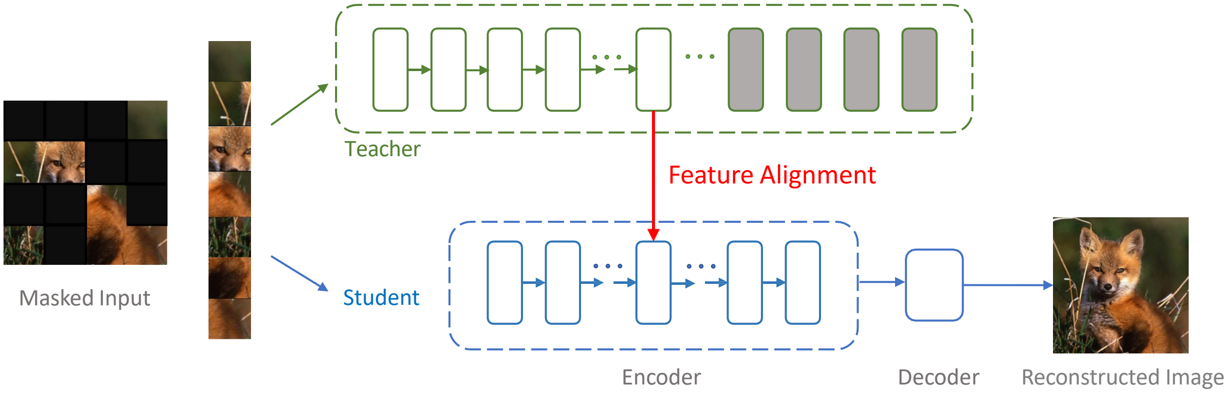

Figure 1. **Illustration of the distillation process in DMAE.** There are two key designs. Firstly, following MAE, we hereby only take visible patches as inputs and aims to reconstruct the masked ones. Secondly, knowledge distillation is achieved by aligning the intermediate features between the teacher model and the student model. Note the gray blocks denote the dropped high-level layers of the teacher model during distillation.

at different capacities. For example, by setting ViT-B as the student model, while the baseline of distilling a fine-tuned ViT-L achieves 82.8% top-1 ImageNet accuracy, DMAE substantially boosts the performance to 84.0% (+1.2%) top-1 ImageNet accuracy, at a even lower training cost (*i.e.*, 195 GPU hours *vs*. 208 GPU hours, see Table 9). More intriguingly, we found that DMAE allows for robust training with extremely highly masked images—even with TEN visible patches (*i.e.*, 95% masking ratio), ViT-B can competitively attain a top-1 ImageNet accuracy of 83.6%; this masking ratio can further be aggressively pushed to 98% (FOUR visible patches) where DMAE still help ViT-B secure 82.4% top-1 ImageNet accuracy. We hope this work can benefit future research on efficiently unleashing the power of pretrain models.

2. Related Work

Knowledge distillation (KD) is a popular model compression technique that allows models to achieve both strong performances of large models and fast inference speed of small models. The first and seminal KD approach, proposed in [17], transfers the "dark knowledge" via minimizing the KL divergence between the soft logits of the teacher model and that of the student model. From then on, many advanced KD methods have been developed, which can be categorized into two branches: logits distillation [8, 12, 21, 30, 34, 35] and intermediate representation distillation [15, 16, 18, 23, 24]. Our DMAE belongs to the second branch, as it minimizes the distance between latent features of the teacher model and those of the student model.

The first feature-based distillation method is FitNets [23]. In addition to aligning logits, FitNets requires the student model to learn an intermediate representation that is predictive of the intermediate representations of the teacher network. Heo et al. [15] re-investigates the design of feature distillation and develops a novel KD method to create a synergy among various aspects, including teacher transform, student transform, distillation feature position, and distance function. CRD [24] incorporates contrastive learning into KD to capture correlations and higher-order output dependencies. Unlike these existing works, our DMAE is the first to consider applying KD to extra information from self-supervised pre-trained models.

Masked image modeling (MIM) helps models to acquire meaningful representations by reconstructing masked images. The pioneering works are built on denoising autoencoders [28] and context encoders [22]. Following the success of BERT in natural language [10], and also with the recent trend of adopting Transformer [27] to computer vision [11], there have emerged a set of promising works on applying MIM for self-supervised visual pre-training. BEiT [3] first successfully adopts MIM to ViT pre-training by learning to predict visual tokens. MaskFeat [29] finds that learning to reconstruct HOG features enables effective visual representation learning. SimMIM [31] and MAE [13] both propose to directly reconstruct the pixel values of the masked image patches. Our work is built on MAE and finds that MAE enables the whole distillation framework to be efficient and effective.

3. Approach

3.1. Masked Autoencoders

Our method is built upon MAE, a powerful autoencoder-based MIM approach. Specifically, the MAE encoder first projects unmasked patches to a latent space, which are then fed into the MAE decoder to help predict pixel values of masked patches. The core elements in MAE include:

Masking. MAE operates on image tokens, *i.e*., the image needs to be divided into non-overlapping patches. A random small subset of those patches will be kept for the MAE encoder, and the rest will be set as the predicting target of the MAE decoder. Typically, a high masking ratio (*e.g*.,75%) is applied, preventing models from taking shortcuts (*e.g*., simply extrapolating missing pixels based on neighbors) in representation learning.

MAE encoder. The MAE encoder is a standard ViT architecture except that it only operates on those unmasked patches. This design largely reduces the computation cost of encoders.

MAE decoder. In addition to the encoded features of unmasked patches (from MAE encoder), the MAE decoder receives mask tokens as input, a learned vector shared across all missing positions. The mask token is only used during pre-training, allowing independent decoder design. Particularly, MAE adopts a lightweight decoder for saving computations.

Reconstruction. Different from BEiT [3] or MaskFeat [29], MAE directly reconstructs image pixel values. The simple mean squared error is applied to masked tokens for calculating loss.

Note that, other than distillation-related operations, the whole pre-training and fine-tuning process in this paper exactly follow the default setup in MAE, unless specifically mentioned. Interestingly, compared to MAE, our DMAE robustly enables a much more efficient pre-training setup, *e.g*., 100 (*vs*. 1600) training epochs and 95% (*vs*. 75%) masking ratio.

3.2. Knowledge Distillation

MAE demonstrates extraordinary capabilities in learning high-capacity models efficiently and effectively. In this work, we seek to combine knowledge distillation with the MAE framework, to efficiently acquire small and fast models with similar performance as those powerful yet cumbersome models. The most straightforward approach is directly applying existing knowledge distillation methods, like the one proposed in DeiT [25], to a fine-tuned MAE model. However, we empirically find that this approach hardly brings in improvements. In addition, this approach fails to leverage the special designs in MAE for reducing computations, *e.g*., only feeding a small portion of the input image to the encoder. To this end, we hereby study an alternative solution: directly applying knowledge distillation at the pre-training stage.

Since there are no categorical labels in MAE pre-training, distilling logits can hardly help learn semantically meaningful representations. We, therefore, resort to distilling the intermediate features. This idea is first developed in FitNets [23], and inspired a set of followups for advancing knowledge distillation [15, 16, 18, 24]. Concretely, we first extract the features from the specific layers of the student model; after feeding such features into a small project head, the outputs will be asked to mimic the features from the corresponding layers of the teacher model. In practice, the projection head is implemented by simple fully connected layer, which addresses the possible feature dimension mismatch between teacher models and student models.

Formally, let $\mathbf{x} \in \mathbb{R}^{3HW\times 1}$ be the input pixel RGB values and $\mathbf{y} \in \mathbb{R}^{3HW\times 1}$ be the predicted pixel values, where H denotes image height and W denotes image width. The MAE reconstruction loss L_{MAE} can be written as

$$L_{MAE} = \frac{1}{\Omega\left(\mathbf{x}_M\right)} \sum_{i\in M} \left(\mathbf{y}_i - \mathbf{x}_i\right)^2. \qquad (1)$$

where M denotes the set of masked pixels, $\Omega(.)$ is the number of elements, and i is the pixel index.

Let $\mathbf{z}_l^S, \mathbf{z}_l^T \in \mathbb{R}^{LC\times 1}$ be the features extracted from the lth layer of the student model and the teacher model, respectively, where L denotes the patch numbers, and C denotes the channel dimension. We use $\sigma()$ to denote the projection network function. Our feature alignment distillation loss L_{Dist} can be written as

$$L_{Dist} = \sum_l \frac{1}{\Omega\left(\mathbf{z}_l^T\right)} \sum_i \left\|\sigma\left(\mathbf{z}_l^S\right)_i - \mathbf{z}_{l,i}^T\right\|_1. \qquad (2)$$

The final loss used in pre-training is a weighted summation of MAE reconstruction loss L_{MAE} and the feature alignment distillation loss L_{Dist}, controlled by the hyper-parameter α:

$$L = L_{MAE} + \alpha \times L_{Dist}. \qquad (3)$$

The framework of DMAE is summarized in Figure 1. Following MAE, DMAE also takes masked inputs and performs the pretext task of masked image modeling. Besides, the corresponding features are aligned between the teacher model and the student model. It is worthy of highlighting that DMAE is an efficient knowledge distiller: 1) it only operates on a tiny subset of visible patches *i.e*., a high masking ratio is applied; and 2) aligning intermediate layer features reduce the computation cost of (cumbersome) teacher model. In the next section, we extensively compare our method with three baselines: the original MAE without any distillation, DeiT-style distillation, and feature alignment distillation in the supervised setting.

4. Experiments

4.1. Implementation Details

Following MAE [13], we first perform self-supervised pre-training on ImageNet-1k [9]. Unless otherwise mentioned, the teacher models are public checkpoints released

# of Layers	Layer Location	Student Aligned Layer Index	ImageNet Top-1 Acc (%)
Single	Bottom	3	82.6
	Middle	6	83.6
		9	**84.0**
	Top	12	83.4
Multiple	Middle+Top	6+12	**84.2**

Table 1. **The effects of feature alignment location.** We hereby test with 5 different layer locations, where top layers refer to those closer to network outputs. For single layer feature alignment, features from the $\frac{3}{4}$ depth of the model leads to the best ImageNet top-1 accuracy. This performance is even comparable to the setup of aligning multiple layers.

from the official MAE implementations[1]. For pre-training, we train all models using AdamW optimizer [20], with a base learning rate of 1.5e-4, weight decay of 0.05, and optimizer momentum $\beta_1, \beta_2 = 0.9, 0.95$. We use a total batch size of 4096, and pre-train models for 100 epochs with a warmup epoch of 20 and a cosine learning rate decay schedule. We by default use the masking ratio of 75%; while the ablation study shows that our method can robustly tackle extremely high masking ratios. *After pre-training, the teacher model is dropped, and we exactly follow the default setups in MAE to fine-tune the student model on ImageNet*.

For feature alignments, we choose to align the features from the $\frac{3}{4}$ depth of both the student model and the teacher model, which we find delivers decent results for all model sizes tested. For example, with a 24-layer ViT-L as the teacher model and a 12-layer ViT-B as the student model, features from the 9th layer of ViT-B are aligned with the features from the 18th layer of ViT-L. We set $\alpha = 1$ in Eq. 3 to balance the tradeoff between MAE reconstruction loss L_{MAE} and the feature alignment distillation loss L_{Dist} in pre-training.

4.2. Analysis

We first provide a detailed analysis of how to set distillation-related parameters in DMAE. Specifically, we set the teacher model as an MAE pre-trained ViT-L (from MAE official GitHub repository, attaining 85.9% top-1 ImageNet accuracy after fine-tuning), and set the student model as a randomly initialized ViT-B. We analyze the following six factors:

Where to align. In Table 1 we first check the effect of feature alignment location on model performance. We observe that shallower features are less favored: *e.g.*, the 3rd layer alignment under-performs all other settings. We speculate this is due to the learning process of ViTs—images are much noisier and less semantic than texts, ViTs will first group the raw pixels in the bottom layers (closer to the input), which is harder to transfer. While features from $\frac{3}{4}$ depth (*i.e.*, the 9th layer in ViT-B) achieves the best performance, a simple rule-of-thumb which we find fits all model scales in our experiments. We adopt this design choice in all other experiments. It is also worth mentioning that simply aligning multiple layers has no clear advantage over aligning features from $\frac{3}{4}$ depth (84.2% *vs*. 84.0%); we, therefore, stick to the $\frac{3}{4}$ depth setting, which is more efficient.

Teacher Layer (relative position)	ImageNet Top-1 Acc (%)
Middle	84.0
Top	83.3
Bottom	82.1

Table 2. **The analysis of aligning order** on ImageNet classification top-1 accuracy (%).

Projection Head	ImageNet Top-1 Acc (%)
Linear	84.0
2-layer MLP	84.0
3-layer MLP	83.8

Table 3. **Projection Head.** A simple fully-connected layer works the best. We choose this as the default setting.

Aligning order. We next test the importance of alignment ordering on model performance. Specifically, by fixing the layer location in the student model (*i.e*., the middle layer in our experiment), we then align it to different layers of the teacher model. As shown in Table 2, we observe that when the aligned layers are in the same relative position (*e.g*., middle to middle), the student model can achieve the best performance.

Masking ratio. MAE reveals that the masking ratio in masked image modeling could be surprisingly high (75%). The hypothesis is that by learning to reason about the gestalt of the missing objects and scenes, which cannot be done by extending lines or textures because of the high masking ratio, the model is also learning useful representations. Interestingly, we find that when combining MAE and knowledge distillation, an even much higher masking ratio is possible, as shown in Figure 2.

Firstly, it is interesting to note that, compared to the typical 75% masking ratio setting, further raising the masking ratio to 90% comes at no performance drop, *i.e*., both attain 84.0% top-1 ImageNet accuracy. Next, even with an extremely large masking ratio like 98% (only FOUR visible patches), DMAE still beats the 100-epoch MAE baseline

[1] https://github.com/facebookresearch/mae

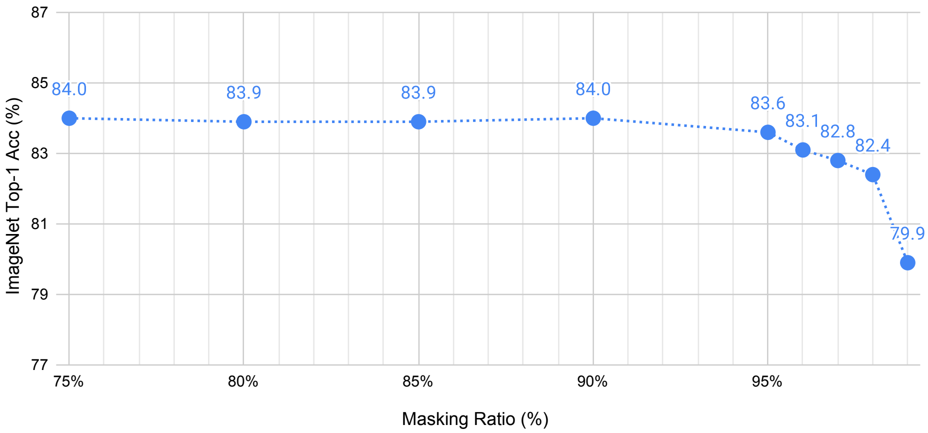

Figure 2. **DMAE allows an extremely high masking ratio.** From left to right, we increase the masking ratio from the basic 75% to the extreme 99%. We note that our DMAE competitively attains 83.6% top-1 ImageNet accuracy with 95% masking ratio (TEN visible patches), and still secures 82.4% top-1 ImageNet accuracy even by learning with FOUR visible patches (98% masking ratio).

Decoder Depth	ImageNet Top-1 Acc (%)
2	83.7
4	83.8
8	84.0

Table 4. **Decoder Depth.** A deeper decoder (slightly) improves the pre-trained representation quality.

	Loss Design	ImageNet Top-1 Acc (%)
Loss Choice	L1 with $\alpha = 1$	84.0
	L2 with $\alpha = 1$	83.3
Loss Ratio	L1 with $\alpha = 0.5$	83.9
	L1 with $\alpha = 1$	84.0
	L1 with $\alpha = 2$	84.0
	L1 with $\alpha = 4$	84.0

Table 5. **Loss Function.** From the first block, we can observe that L1 distance yields significantly higher performance than L2 distance. From the second block, we can observe that the hyper-parameter α has little influence on the representation quality of DMAE.

that uses a masking ratio of 75% (second row in Table 6), by a non-trivial-margin (82.4% *vs*. 81.6%). These results suggest that, with the assistance of distilled knowledge from the teacher model, the student model can make better use of visible patches, even at a very limited amount, for representation learning.

Projection head. The goal of the proposed feature alignment distillation is to encourage the student model to learn features that are predictive of features from a stronger teacher model. To that end, a small projection head is employed on features from the student model, to 1) project them onto a space of the same dimension as the hidden dimension of the teacher model, and 2) provide extra flexibility for feature alignment. We ablate the choice of this projection network, as shown in Table 3. We can observe that applying a simple fully connected layer already performs the best among other choices.

Decoder depth. In Table 4 we analyze the effect of decoder depth. Similar to MAE [13], the final performance gets (slightly) increased with a deeper decoder. We choose a decoder depth of 8 as the default setting as in [13]. Note that a decoder depth of 2 is also a competitive choice—compared to an 8-depth decoder, it significantly reduces the computation cost while only marginally sacrificing the accuracy by 0.3%.

Loss designs. Table 5 ablates the loss design. While SimMIM [31] shows that L1 distance and L2 distance lead to similar performance, ours suggests that L1 distance exhibits a clear advantage over L2 distance, *i.e.*, +0.7% improvement. Furthermore, we note DMAE is quite robust to the specific value of the hyperparameter α, which controls the relative importance of the distillation loss over the reconstruction loss. Based on these results, we choose L1 in Eq. 2 for distance measure and set $\alpha = 1$ in Eq. 3 for the rest experiments.

4.3. Comparison with Baselines.

In Table 6, we compare the performance of our DMAE with various baselines:

MAE. The MAE baselines are presented in the first block of Table 6. If MAE is also asked to pre-train for only 100 epochs, DMAE can substantially outperform this baseline by 2.4% (from 81.6% to 84.0%). When comparing to a much stronger but more computationally expensive MAE baseline with 1600 pre-training epoch, we note DMAE still beats it by 0.4%.

Supervised model. The second block in Table 6 demonstrates the effectiveness of DMAE compared with models

Method	Pre-training epochs	Supervised training / fine-tuning epochs	ImageNet Top-1 Acc (%)
MAE-B	100	100	81.6
MAE-B	1600	100	83.6
DeiT-B	-	100	76.8
DeiT-B	-	300	81.8
DeiT-B-Soft Distillation	-	100	77.5
DeiT-B-Hard Distillation	-	100	78.3
CRD [24]	-	100	81.9
SRRL [32]	-	100	82.2
Dear-KD [6]	-	100	82.4
Supervised Feature Alignment	-	100	82.8
DMAE-B	100	100	84.0

Table 6. **DMAE shows stronger performance than all three kinds of baselines**: MAE, supervised model, and other existing advanced distillation strategies.

trained under the supervision of categorical labels, which requires a much longer training time, *i.e*., +2.2% compared to the DeiT 300 epochs supervised training.

Other distillation strategies. We next compare DMAE with other distillation methods. We consider DeiT-style logit-based distillation [25], CRD [24], SRRL [32], Dear-KD [6], and feature alignment distillation. Note that for these baselines, one significant difference from DMAE is that the student model here directly distill knowledge from a supervisely fine-tuned teacher model. Moreover, to make these baselines more competitive, the teacher models will first be MAE pre-trained and then fine-tuned on ImageNet-1k.

The results are shown in the third block of Table 6. Firstly, we can observe that the DeiT-style logit-based distillation, either soft or hard, even hurts the student models' performance. This phenomenon potentially suggests that such a distillation strategy may not fit teacher models of ViT architectures. For other baselines, we note that feature alignment distillation performs the best; but this is still worse than DMAE (82.8% *vs*. 84.0%), indicating the importance and effectiveness of distilling knowledge from a pre-trained teacher model.

4.4. Scaling to Different Model Sizes

We test DMAE with different model sizes, listed in Table 7. For a fair comparison, both methods only pre-train models for 100 epochs. DMAE shows consistent improvement compared to MAE across different model sizes. With *only one middle layer* feature alignment, DMAE brings an additional improvement of +2.4% with ViT-B, +2.7% with ViT-Small, and +3.4% with ViT-Tiny. In addition, we are interested in the following two cases:

Same teacher model, different student model. As shown in the first two lines in Table 7, we find that when using ViT-L as the teacher model, both ViT-B and ViT-S benefit from the distillation, demonstrating a clear advantage over the MAE baseline. We argue that the ability to effectively generalize to cases where an even smaller student model is desirable, especially for those computation-constrained real-world applications.

Student Model	Teacher Model	ImageNet Top-1 Acc (%)	
		MAE	DMAE
Base	Large	81.6	84.0 (+2.4)
Small	Large	77.4	80.1 (+2.7)
	Base	77.4	79.3 (+1.9)
Tiny	Base	66.6	70.0 (+3.4)

Table 7. **Across different model sizes**, DMAE shows consistent improvements compared with MAE.

Same student model, different teacher model. As shown in Table 7, with a ViT-S as the student model, enlarging the teacher model from ViT-B to ViT-L further boosts the accuracy by 0.8% (from 79.3% to 80.1%). This result suggests that DMAE can effectively distill knowledge from the teacher models at different scales.

4.5. Limited Training Data

In certain real-world applications, data could be hard to acquire because of high data collection and labeling costs or due to privacy concerns. Leveraging models pre-trained on large-scale unlabeled datasets for fine-tuning when only a small dataset of downstream tasks is available becomes a promising solution. Here we seek to test the potential of our DMAE in this data-scarce scenario. We strictly follow [5] to sample 1% or 10% of the labeled ILSVRC-12 training datasets in a class-balanced way. We set MAE pre-trained ViT-L as the teacher model and a randomly initialized ViT-B as the student model. In addition, we compare DMAE with three kinds of baselines described in Section 4.3: MAE, supervised model, and other distillation strategies, and similarly, set the teacher model to be a ViT-L that first pre-trained with MAE and then fine-

Method	Pre-training epochs	Supervised training / fine-tuning epochs	IN-1% Top-1 Acc(%)	IN-10% Top-1 Acc(%)
MAE-B	100	100	33.9	65.0
MAE-B	1600	100	49.6	72.8
DeiT-B	-	100	-	-
DeiT-B-Soft Distillation	1600	100	36.0	66.4
DeiT-B-Hard Distillation	1600	100	37.3	67.3
Supervised Feature Alignment	1600	100	34.2	67.6
DMAE-B	100	100	50.3	73.4

Table 8. **DMAE demonstrates much stronger performance than all other baselines when training data is limited.** Note that the DeiT-B baseline is unable to converge because of data insufficiency.

Model	Training Cost (GPU Hours)			ImageNet
	Pre-training	Fine-tuning	Overall	Top-1 Acc(%)
MAE-B-100-epoch	78h	112h	190h	81.6
MAE-B-1600-epoch	1248h	112h	1360h	83.6
DeiT-B-Soft Distillation	-	213h	213h	77.5
DeiT-B-Hard Distillation	-	213h	213h	78.3
Supervised Feature Alignment	-	208h	208h	82.8
DMAE-B	83h	112h	195h	**84.0**

Table 9. **Computational cost comparisons among DMAE and other baselines.** The training cost is measured by A5000 GPU hours. We note the proposed DMAE maintains a similar (or even cheaper) training cost than others, while achieving much higher top-1 ImageNet accuracy.

tuned on ImageNet-1k. Note that since DMAE has full access to the 100% ImageNet dataset (without labels) during pre-training, to ensure a fair and competitive comparison, *we initialize all the baselines as the 1600-epoch MAE pre-trained model on ImageNet*.

Table 8 shows that DMAE largely surpasses all other baselines. For example, when only 10% ImageNet data is available for supervised training or fine-tuning, DMAE outperforms the MAE pre-trained baseline by 8.4% (*i.e*., 73.4% *vs*. 65.0%). DMAE also significantly outperforms other distillation strategies, with an improvement ranging from 5.8% to 7.0%. We note this accuracy gap is even larger when only 1% ImageNet is available, demonstrating the data efficiency of DMAE.

4.6. Generalization to Other Methods

Lastly, we provide preliminary results of integrating DMAE into other self-supervised training frameworks, including DINO [4] and MoCo-V3 [7]. Unlike MAE, which belongs to masked image modeling, DINO and MoCo-V3 are contrastive learning-based methods. Still, as shown in Table 10, without further hyperparameter tuning, DMAE effectively shows non-trivial improvements on top of both DINO (+1.3%) and MoCo-v3 (+1.4%), demonstrating the potential of distilling pre-trained models (rather than fine-tuned models as in most existing knowledge distillation frameworks).

	w/o DMAE	w/ DMAE
DINO	80.9	**82.2 (+1.3)**
MoCo-V3	81.1	**82.5 (+1.4)**

Table 10. **DMAE effectively improves other self-supervised pre-training frameworks** (including DINO and MoCo-v3) on ImageNet classification top-1 accuracy (%).

4.7. Generalization to Downstream tasks

Following ViT-Det [19], we conduct downstream fine-tuning on the COCO dataset, where MAE/DMAE pre-trained ViT-Base model is adopted as the plain backbone of Mask-RCNN. We train the model on the `train2017` split and evaluate it on the `val2017` split. We report results on bounding box object detection (AP^{box}) and instance segmentation (AP^{mask}), shown in Table 11. Our observations indicate that DMAE also demonstrates strong potential in downstream tasks like detection and segmentation.

method	AP^{box}	AP^{mask}
MAE	51.2	45.5
DMAE	**53.4**	**46.9**

Table 11. **DMAE effectively improves the downstream tasks** as well, including object detection and segmentation.

5. Discussion

In this section, we present a quantitative evaluation of the computational cost of DMAE compared to baseline methods. Additionally, we conduct a standard deviation analysis to assess the stability of our DMAE approach. Finally, we propose an advanced fine-tuning recipe for applying DMAE to smaller ViT models.

5.1. Computational Costs

In Table 9, we provide a quantitative evaluation on the computational cost, which is tested on a single NVIDIA A5000 GPU. We can observe that, without a significantly increasing of training hours, DMAE substantially outperforms the MAE-100 baseline by +2.4% on ImageNet; this result even exceeds the MAE-1600 baseline, which causes ~7x GPU hours than MAE-100 baseline. We additionally provide the actual GPU hours of other baselines, and find that the proposed DMAE stands as the most efficient one, meanwhile achieving the best top-1 ImageNet accuracy.

5.2. Standard Deviation Analysis

In the above experiments, we kept the same random seed. Following MAE, we perform the statistical analysis for DMAE by changing the random seeds. In Table 12, from top to down, we show three aligning settings; and from left to right, we show the results with the default seed, the average accuracy with three randomly sampled seeds, and their standard deviation, respectively. From these results, we could conclude that our DMAE can bring in statistically stable improvements.

Pos	Acc(%)	Avg	Standard Deviation
Bottom (3th)	82.6	82.50	0.20
Mid (6th)	83.6	83.67	0.08
Top (9th)	84.0	84.03	0.10

Table 12. **Standard deviation analysis for DMAE.** From top to bottom, we show three aligning settings, the model performance with the default seed, the average performance with three randomly sampled seeds, and their standard deviations, respectively. ’Pos’ denotes the student distillation position. Acc denotes ImageNet Top-1 accuracy. ’Avg’ denotes the average value over three times.

5.3. Smaller ViT Models.

Since the original MAE paper does not offer specialized recipes for ViT-Small and Tiny, we by default use the recipe for ViT-Base to fine-tune these smaller ViTs. However, this recipe includes strong regularization and augmentation techniques that might lead to over-regularization for the smaller ViTs. To address this issue, we experiment with a modified recipe with weaker augmentation and regularization by removing MixUp, CutMix, and Stochatic Depth. Results on ImageNet in Table 13 show that DMAE not only maintains its advantage over MAE, but also outperforms DeiT with this new recipe, highlighting its effectiveness.

ViT-Tiny	Top-1 Acc (%)
MAE	70.1
DeiT	74.5
DMAE	**74.9**

(a) ViT-Tiny

ViT-Small	Top-1 Acc (%)
MAE	80.0
DeiT	81.2
DMAE	**82.2**

(b) ViT-Small

Table 13. **Weaker augmentation and regularization helps smaller ViT models**, during finetuning for both MAE and DMAE on ImageNet classification top-1 accuracy (%)

6. Conclusion

Self-supervised pre-training has demonstrated great success for those exponentially growing models in the natural language domain. Recently, the rise of MAE shows that a similar paradigm also works for the computer vision domain, and now the development of vision models may embark on a similar trajectory as in the language domain. Yet, it is often desirable to have a well-balanced model between performance and speed in real-world applications. This work is a small step towards unleashing the potential of knowledge distillation, a popular model compression technique, within the MAE framework. Our DMAE is a simple, efficient, and effective knowledge distillation method: feature alignment during MAE pre-training. Extensive experiments on multiple model scales demonstrate the effectiveness of our approach. Moreover, an intriguing finding is that it allows for a masking ratio even higher than the already large one used in MAE (*i.e*., 75%). We have also validated the effectiveness of our DMAE when in the small-data regime. We hope this work can benefit future research in knowledge distillation with pre-trained models.

Acknowledgement

This work is supported by a gift from Open Philanthropy, TPU Research Cloud (TRC) program, and Google Cloud Research Credits program. We also want to acknowledge ONR with award N00014-21-1-2812.

References

[1] Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 16061–16070, 2022. 1

[2] Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020. 1

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 1, 2, 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021. 7

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 6

[6] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *CVPR*, pages 12052–12062, 2022. 6

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 7

[8] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *CVPR*, pages 4794–4802, 2019. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2

[12] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616. PMLR, 2018. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1, 2, 3, 5

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[15] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *CVPR*, pages 1921–1930, 2019. 2, 3

[16] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, volume 33, pages 3779–3787, 2019. 2, 3

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2

[18] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *NeurIPS*, 31, 2018. 2, 3

[19] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022. 7

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[21] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, volume 34, pages 5191–5198, 2020. 1, 2

[22] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2

[23] 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. 2, 3

[24] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 2, 3, 6

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 1, 3, 6

[26] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *CVPR*, pages 32–42, 2021. 1

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 2

[28] 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. *JMLR*, 11(12), 2010. 2

[29] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14668–14678, 2022. 2, 3

[30] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3252–3262, 2022. 2

[31] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2, 5

[32] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021. 6

[33] Qihang Yu, Yingda Xia, Yutong Bai, Yongyi Lu, Alan L Yuille, and Wei Shen. Glance-and-gaze vision transformer. *Advances in Neural Information Processing Systems*, 34:12992–13003, 2021. 1

[34] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 1, 2

[35] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. *arXiv preprint arXiv:2203.08679*, 2022. 2