On Data Scaling in Masked Image Modeling

Zhenda Xie[13], Zheng Zhang[3†], Yue Cao[3†], Yutong Lin[23], Yixuan Wei[13], Qi Dai[3], Han Hu[3]
[1]Tsinghua University [2]Xi'an Jiaotong University [3]Microsoft Research Asia
{t-zhxie, zhez, yuecao, t-yutonglin, t-yixuanwei, qi.dai, hanhu}@microsoft.com

Abstract

Scaling properties have been one of the central issues in self-supervised pre-training, especially the data scalability, which has successfully motivated the large-scale self-supervised pre-trained language models and endowed them with significant modeling capabilities. However, scaling properties seem to be unintentionally neglected in the recent trending studies on masked image modeling (MIM), and some arguments even suggest that MIM cannot benefit from large-scale data. In this work, we try to break down these preconceptions and systematically study the scaling behaviors of MIM through extensive experiments, with data ranging from 10% of ImageNet-1K to full ImageNet-22K, model parameters ranging from 49-million to one-billion, and training length ranging from 125K to 500K iterations. And our main findings can be summarized in two folds: 1) masked image modeling remains demanding large-scale data in order to scale up computes and model parameters; 2) masked image modeling cannot benefit from more data under a non-overfitting scenario, which diverges from the previous observations in self-supervised pre-trained language models or supervised pre-trained vision models. In addition, we reveal several intriguing properties in MIM, such as high sample efficiency in large MIM models and strong correlation between pre-training validation loss and transfer performance. We hope that our findings could deepen the understanding of masked image modeling and facilitate future developments on large-scale vision models. Code and models will be available at https://github.com/microsoft/SimMIM.

1. Introduction

Masked Image Modeling (MIM) [3, 18, 44], which has recently emerged in the field of self-supervised visual pre-training, has attracted widespread interest and extensive applications throughout the community for unleashing the superior modeling capacity of attention-based Transformer architectures [13, 26] and demonstrating excellent sample efficiency and impressive transfer performance on a variety of vision tasks. However, most recent practices are focused on the design of MIM methods, while the study of scaling properties of MIM is unintentionally neglected, especially the data scaling property, which successfully motivated the large-scale self-supervised pre-trained language models and endowed them with significant modeling capabilities. Although previous works [8, 10, 17, 37, 45] have explored several conclusions about the scaling properties of vision models, most of their findings were obtained under a supervised pre-training scheme or under a contrastive learning framework, so the extent to which these findings could be transferred to MIM still needs to be investigated.

Meanwhile, with the emergence of Transformers [41] and masked language modeling (MLM) [12, 31], the systematic studies of scaling laws have already been explored in natural language processing field [21, 23, 35], which provided ample guidance for large models in recent years. The core finding drawn from scaling laws [21] for neural language models is that *the performance has a power-law relationship with each of the three scale factors – model parameters N, size of dataset D, and amount of compute C respectively – when not bottlenecked by the other two*. This conclusion implies that better performance can be obtained by scaling up these three factors to the extent that the scaling laws are in effect, which led to the subsequent developments of large scale language models [16, 28, 29, 32, 33] that exhibit excellent modeling capabilities [4] on most language tasks. Therefore, it is natural to ask whether MIM possesses the same scaling signatures for vision models as the MLM method for language models, so that the scaling up of vision models can catch up with language models.

Though masked image modeling and masked language modeling both belong to masked signal prediction, their property differences are also non-negligible due to the different nature of vision and language. That is, the images are highly redundant raw signals and words/sentences are semantically rich tokens, which may result in different abilities of data utilization, and it is thus debatable whether the observations from language models could be reproduced in vision models.

The work is done when Zhenda Xie, Yutong Lin, and Yixuan Wei are interns at Microsoft Research Asia. † Project co-leaders.

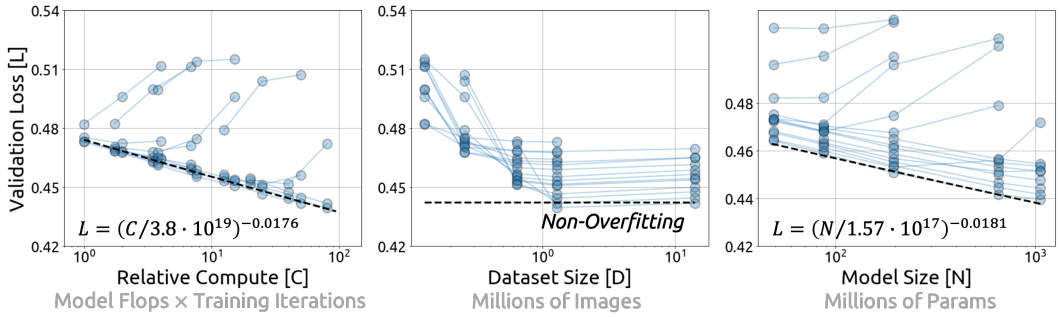

Figure 1. The curves of validation loss of pre-training models w.r.t. the relative compute, dataset size and model size. MIM performance improves smoothly as we increase the relative compute and model size, but not improve when the dataset size is sufficient to prevent model from overfitting. We set the relative compute of SwinV2-S for 125K iterations as the value of 1. *Best viewed in color.*

Moreover, recent studies [14, 38] have shown that using a small amount of training data in masked image modeling can achieve comparable performance to using large datasets, which also motivates our explorations as it disagrees with previous findings and intuitions.

In this paper, we systematically investigate the scaling properties, especially the data scaling capability of masked image modeling in terms of different dataset sizes ranging from 10% of ImageNet ($\sim$ 0.1 million) to full ImageNet-22K ($\sim$ 14 million), model sizes ranging from 49-million Swin-V2 Small to 1-billion Swin-V2 giant and training lengths ranging from 125K to 500K iterations. We use Swin Transformer V2 [25] as the vision encoder for its proven trainability of large models and applicability to a wide range of vision tasks, and adopt SimMIM [44] for masked image modeling pre-training because it has no restrictions on encoder architectures. We also conduct experiments with other MIM frameworks like MAE [18] and other vision encoder like the widely used ViT [13] to verify the generalizability of our findings. With these experimental setups, our main findings could be summarized into two folds:

i) *Masked image modeling remains demanding large-scale data in order to scale up computes and model parameters.* We empirically find that MIM performance has a power-law relationship with relative compute and model size when not bottlenecked by the dataset size (Figure 1). Besides, we observe that smaller datasets lead to severe overfitting phenomenon for training large models (Figure 2), and the size of the dataset to prevent model from overfitting increases clearly as the model increases (Figure 5-Left). Therefore, from the perspective of scaling up model, MIM still demands large-scale data. Furthermore, if we train large-scale models of different lengths, we find that relatively small datasets are adequate at shorter training lengths, but still suffer from overfitting at longer training lengths(Figure 5-Right), which further demonstrate the data scalability of MIM for scaling up compute.

2) *Masked image modeling cannot benefit from more data under a non-overfitting scenario, which diverges from the previous observations in self-supervised pre-traind language models or supervised pre-trained vision models.* In MIM, we find that increasing the number of unique samples for a non-overfitting model does not provide additional benefits to performance (Figure 1). This behavior differs from previous observations in supervised vision transformers and self-supervised language models, where the model performance increases as the number of unique samples increases.

In addition, we also demonstrate some intriguing properties about masked image modeling, such as larger models possessing higher sample efficiency, i.e., fewer optimization steps are required for larger models to achieve same performance (Section 4.2); and the consistency between transfer and test performance, i.e., test performance could be used to indicate the results on downstream tasks (Section 4.3). These observations also correspond to those in previous practices.

These findings on the one hand confirm the effect of MIM on scaling up model size and compute, and raise new concerns and challenges on the data scalability of MIM on the other. We hope that our findings could deepen the understanding of masked image modeling and facilitate future developments on large-scale vision models.

2. Related Work

Masked Image Modeling Masked Image Modeling learns representations by reconstructing the masked content of images, and its early exploration can be traced back to context encoder [30] and denoising autoencoder [42]. Recently, iGPT [6], BEiT [3], MAE [18] and SimMIM [44] recall this approach on training vision transformer. iGPT [6] sequentially predicted the pixels by auto-regressive manner. BEiT [3] proposed to predict the discrete visual tokens. MAE [18] and SimMIM [44] concurrently found predicting the raw pixels with a high masking ratio could work well. In this work, we use SimMIM as the default masked image modeling approach, because of its simplicity and no restrictions on the architecture of vision encoder like MAE.

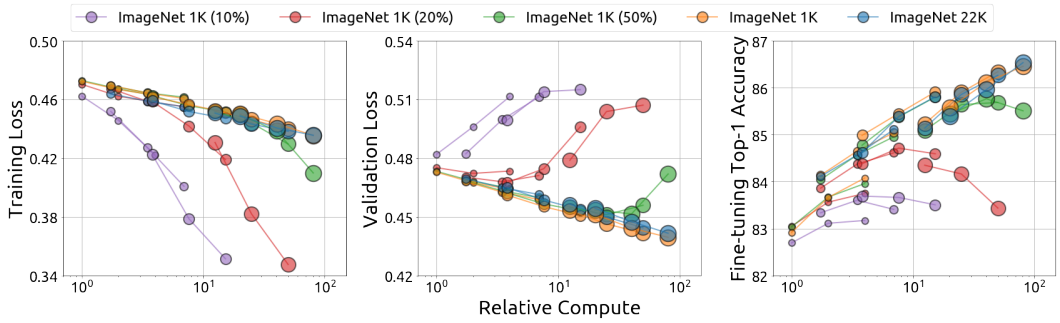

Figure 2. Training loss of pre-training, validation loss of pre-training, and fine-tuning accuracy on ImageNet-1K of different model sizes, data sizes and training lengths, w.r.t. the relative pre-training compute. All models are SwinV2 models pre-trained with SimMIM. We set the training compute of SwinV2-S for 125K iterations as the value of 1. Bigger circles indicate larger models. *Best viewed in color.*

Vision Transformers Transformer [41] was first applied to natural language processing and became the dominant architecture, and has recently attracted a lot of attention in computer vision. The pioneering work ViT [13] first shows that the transformer architecture works well in image classification when trained on large amounts of data. DeiT [39] proposed a better training recipe based on ViT and demonstrated that vision Transformer has promising performance when only using ImageNet-1K dataset. Swin Transformer [26] improves plain ViT by inducing the hierarchical architecture and non-overlapping local attention and successfully demonstrates the effectiveness of vision transformer on a wide range of vision tasks. Swin Transformer V2 [25] further addresses the training stability issue of [26] in model scaling and illustrates better performance than the original Swin Transformer, and thus we use it as the default vision encoder in this work.

Scaling Vision Models Many works [25, 34, 37, 45] examine how to scale vision models, but most are more concerned with exploring the perspective of model architecture designs. For example, EfficientNet [37] extensively studied how model width, model depth and input resolution affect the convolutional neural networks; [34] proposed to scale vision model with sparse mixture-of-expert; [45] and [25] studied how to scale ViT and Swin Transformer, respectively. In addition, several studies [1, 2, 45] explored the aspect of data scaling under the pre-training fine-tuning paradigm. BiT [22] revisited the supervised pre-training on a wide range of data scales up to 1M images. SEER [17] studied the effectiveness of data scaling in the contrastive learning framework with up to one billion images. Recently, SplitMask [14] find that masked image modeling is robust to the size of pre-training data and challenges the data scaling capability of masked image modeling, which is most relevant to our work.

Scaling Language Models Many works [20, 23, 35] systematically analyzed the scaling behavior on a relatively small scale. And [21] studied empirical scaling laws for large scale neural language models and revealed that the loss scaled as a power-law with model size, dataset size, and the amount of compute used for training. Motivated by these laws, several large scale language models [4,16,28,29,32,33] were successfully trained and demonstrated excellent modeling capabilities on most language tasks. [36] established a comprehensive, large-scale benchmark designed to assist in quantifying and extrapolating the capabilities of large language models.

3. Background and Experimental Setup

3.1. Masked Image Modeling

Masked image modeling is used to train the vision model by taking a corrupted image as input and predicting the content of the masked region as the target. In this study, we use SimMIM [44] as the default masked image modeling approach because of its simplicity and lack of restrictions on the architecture of the vision encoder. SimMIM consists of a visual encoder and an extremely lightweight prediction head of a linear layer for predicting the raw pixels of the corrupted images via ℓ_2 regression loss. To facilitate the implementation of the vision transformer, SimMIM adopts the patch-wise mask strategy with the masked patch size of 32×32 and mask ratio of 0.6. To further alleviate the local dependency of raw pixels, we improved the SimMIM by normalizing the predicted target according to [15] with a sliding window of 47^2. As the result, a slight performance improvement is observed. In addition, we conduct experiments using MAE as the masked image modeling approach to verify the methodological generalizability of our findings. These experiments strictly follow the settings in [18].

3.2. Architecture Specifications

We use Swin Transformer V2 [25] as the default vision encoder in this study. Thanks to its generality and scalability, we evaluate a series of SwinV2 models with a wide range of model sizes (the number of parameters ranges from $\sim$50M to $\sim$1B, and FLOPs range from $\sim$9G to $\sim$190G) on

Model	Base Channel	Depth	Head	Window Size		Backbone Params
				pre-train	fine-tune	
SwinV2-S	96	{2, 2, 18, 2}	{3, 6, 12, 24}	12	14	49M
SwinV2-B	128	{2, 2, 18, 2}	{4, 8, 16, 32}	12	14	87M
SwinV2-L	192	{2, 2, 18, 2}	{6, 12, 24, 48}	12	14	195M
SwinV2-H	352	{2, 2, 18, 2}	{11, 22, 44, 88}	12	14	655M
SwinV2-g	448	{2, 2, 18, 2}	{14, 28, 56, 112}	12	14	1061M

Table 1. Detailed architecture specifications. Note that, the model aliases used in our work are different from [25].

	IN1K (10%)	IN1K (20%)	IN1K (50%)	IN100	IN1K(100%)	IN22K(100%)
#Classes	1×10^3	1×10^3	1×10^3	1×10^2	1×10^3	2.18×10^4
#Images	1.28×10^5	2.56×10^5	6.41×10^5	1.27×10^5	1.28×10^6	1.42×10^7

Table 2. Detailed dataset specifications used in the pre-training of masked image modeling.

multiple downstream tasks. The detailed model specifications are shown in Table 1. We use a new variant SwinV2-g (giant), with number of parameters between SwinV2-L and the 3-billion-parameter SwinV2-G (Giant) used in [25]. In addition, we conduct experiments with a series of ViT models [13] to prove the architectural generalizability of our findings. Specifically, we use ViT-B/16, ViT-L/16 and ViT-H/14 according to the settings from [18].

3.3. Pre-training Datasets

To study the effect of data size on masked image modeling, we build datasets with different sizes. We use the training set of ImageNet-1K and ImageNet-22K as two large-scale datasets, and randomly sample 10%, 20%, 50% of images in the ImageNet-1K training set as smaller datasets. By default, the images are uniformly sampled from each category. We also consider the sampling strategies could perform differently. To this end, we randomly sample 100 classes from ImageNet-1K as ImageNet-100, and compare it with ImageNet-1K (10%) but find their training loss and fine-tuning performance are almost the same. The details and statistics of all pre-training datasets used in our study are shown in Table 2.

3.4. Pre-training Details

To better compare the performance of models with different amounts of data under the same pre-training length, we use training iterations rather than training epochs and adopt the same hyper-parameters for all models with different sizes during pre-training. The total number of training iterations is in {125K, 250K, 500K} and the batch size is set as 2048 for all experiments. In pre-training stage, we use the same hyper-parameters for all models, and the training details and hyper-parameters of pre-training are summarized in Appendix. Because of the excessive amount of experiments, we follow SimMIM [44] and also use the following two techniques for reducing the experimental overheads: First, we use the step learning rate scheduler in pre-training for sharing the first training step among experiments with different training lengths. The first 7/8 training iterations are the first step and the last 1/8 training iterations are the second step with the learning rate ratio of 0.1 (*i.e.* learning rate is divided by 10 in the second step). Second, we adopt the input image size of 192^2 and set the window size of 12. We improve the SimMIM by normalizing the predicted target according to [15] with a sliding window of 47^2 and observe an improvement of 0.3 on top-1 accuracy of ImageNet-1K for the SwinV2-Large model. The same light data augmentation strategy as SimMIM is used: random resize cropping with a scale range of [0.67, 1], an aspect ratio range of [3/4, 4/3] and a random flipping with probability 0.5.

3.5. Fine-tuning Tasks

To extensively and accurately evaluate the performance of pre-trained models under different pre-training schedulers and datasets, a series of diverse and representative tasks including fine-tuning on ImageNet-1K, fine-grained image classification, object detection, instance segmentation, and semantic segmentation are selected for evaluation. Detailed setups of all experiments are illustrated in Appendix.

ImageNet-1K We follow [3] to evaluate the quality of learnt representations by fine-tuning the pre-trained models on ImageNet-1K [11] image classification task, which is the most commonly used scenario and evaluation criterion for pre-trained models [18, 44]. Different from pre-training, We adopt the image size with 224^2 with window size of 14 in fine-tuning. The AdamW with batch size of 2048, base learning rate of 5e-3, weight decay of 0.05, β_1 of 0.9 and β_2 of 0.999 are used, and we adopt cosine learning rate scheduler. As larger models are more prone to overfitting, we fine-tune SwinV2-S/B/L for 100 epochs with 20 warm-up epochs and SwinV2-H/g for 50 epochs with 10 warm-up epochs, and decrease the layer decay as the model size increases. In addition, gradient clipping, stochastic depth, label smoothing and data augmentations (*e.g.* random crop,

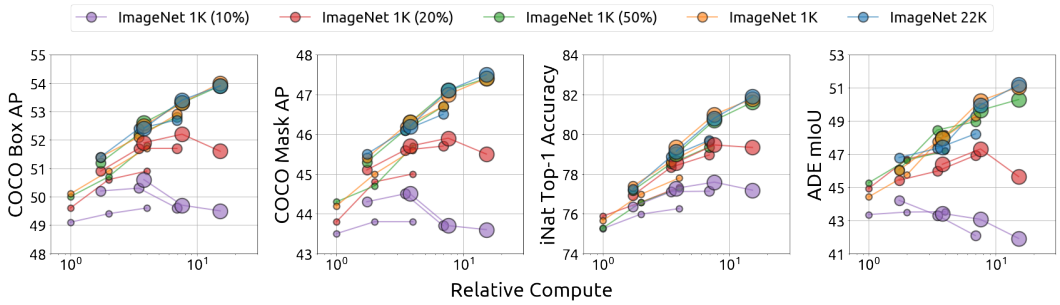

Figure 3. **Left to right:** Performances with SwinV2 models on (a) COCO object detection, (b) COCO instance segmentation, (c) iNaturalist-18, and (d) ADE20K semantic segmentation w.r.t. the relative compute. Note that the relative compute indicates the pre-training compute. We set the relative compute of SwinV2-S for 125K iterations as 1. Bigger circles indicate larger models. *Best viewed in color.*

rand erasing [47], rand augment [9], mixup [46], cutmix [46], *etc.*) are also used by following [44].

iNaturalist-18 iNaturalist [40] 2018 is a long-tailed fine-grained image classification dataset. As fine-tuning in ImageNet-1K, we also use the input image size of 224^2, window size of 14 and patch size of 4 in iNaturalist 2018. We fine-tune all models for 100 epochs with 20 warm-up epochs, and set layer decay to 0.8, 0.75 and 0.7 for SwinV2-S/B/L, respectively. The AdamW optimizer with cosine learning rate scheduler, batch size of 2048, base learning rate of 1.6e-2, weight decay of 0.1, β_1 of 0.9 and β_2 of 0.999 are used. In addition, we also adopt stochastic depth, label smoothing, gradient clipping and data augmentations in fine-tuning.

COCO Object Detection and Instance Segmentation [24] We use Mask R-CNN [19][1] for evaluation. We set the window size to 14 and patch size to 4. The AdamW optimizer with batch size of 32, base learning rate of 8e-5, weight decay of 0.05, β_1 of 0.9, β_2 of 0.999 and a step learning rate scheduler (step learning rate ratio of 0.1, step epochs are 27 and 33) are used. In training, the random cropping with crop size of [1024, 1024], large scale jittering with a range of [0.1, 2.0], random horizontal flip with probability 0.5, and stochastic depth regularization are used. In testing, all images are resized to (800, 1333) and keeping the aspect ratio unchanged.

ADE20K Semantic Segmentation [48] Following [26], we use UPerNet [43] for evaluation. We set the window size to 20 and the patch size to 4. The AdamW optimizer with with batch size of 32, base learning rate searched in a range of [1e-4, 3e-4], weight decay of 0.05, β_1 of 0.9, β_2 of 0.999 and a linear learning rate scheduler with a total of 80K iterations are used. Also, we use the layer decay of 0.95, 0.95, 0.9 for SwinV2-S/B/L, respectively. In training, the random cropping with crop size of [640, 640], scale jittering with a range of [0.5, 2.0], random horizontal flip with probability 0.5, random photometric distortion and stochastic depth regularization of 0.1 are used. In testing, all images are evaluated by sliding window manner, and use the test image size of (2560, 640) and set sliding window stride to 426, following [26, 44].

[1]Our implementation based on MMDetection [5].

4. Results and Findings

We train numerous models with different training lengths, dataset sizes, and model sizes, and study how these factors affect the performance of masked image modeling. Figure 1 illustrates the validation loss of pre-training[2] with respect to the relative compute, dataset size and model size. Figure 2 illustrates the training loss of pre-training, validation loss of pre-training, and the fine-tuning top-1 accuracy of ImageNet-1K with respect to the relative compute. Based on these extensive experiments, we make the following observations:

4.1. Data Scaling in Masked Image Modeling

Masked image modeling remains demanding for large-scale data. When with the high masking rate (*e.g.*, 60% in our work), the masked image modeling is considered a very challenging training objective and has been found to be data efficient by previous literature [14,27], *i.e.*, a comparable performance can be achieved with small datasets as with large datasets. However, Figure 2 shows that as the training cost increases, the training loss of some models drops significantly, and their validation loss rises significantly, even on using 50% images of ImageNet-1K (*i.e.*, IN1K (50%)), indicating the *overfitting* phenomenon exists. In Figure 5-Left, we demonstrate that the size of the dataset to prevent model from overfitting increases clearly as the model increases. And significant decrease to the fine-tuning performance caused by overfitting could be observed in Figure 2 and Figure 3. Moreover, we measure the best fine-tuning performance of each model trained by different training schedulers in Figure 2-Right. We find the large models perform even worse than smaller models when small dataset is used for training. For

[2]The validation loss of pre-training is measured on the validation set of ImageNet-1K for all experiments.

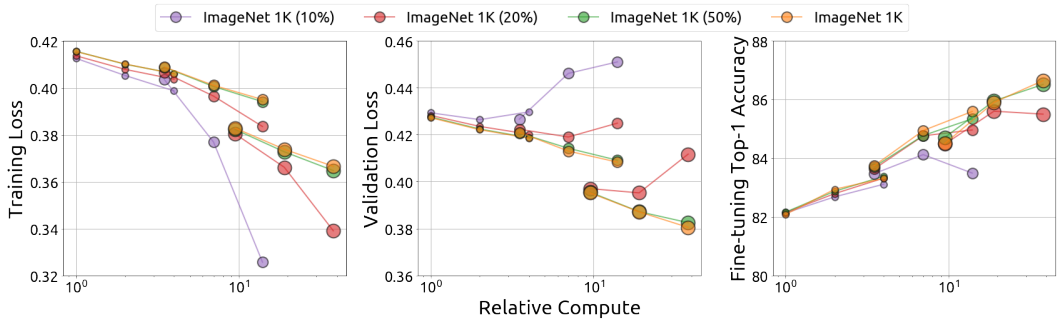

Figure 4. Training loss of pre-training, validation loss of pre-training, and fine-tuning accuracy on ImageNet-1K of different model sizes, data sizes and training lengths, w.r.t. the relative pre-training compute. All models are ViT models pre-trained with MAE. We set the training compute of ViT-B/16 for 125K iterations as the value of 1. Bigger circles indicate larger models. *Best viewed in color.*

example, the best top-1 accuracy of SwinV2-H with IN1K (20%) is 84.4, worse than the best performance of SwinV2-L by 0.3. In addition, by comparing the best performance that can be obtained using different sizes of dataset, we find that using more data results in better performance. These observations suggest that masked image modeling does not alleviate the demands of large dataset.

The training length matters. Larger models can benefit from more data at a longer training length. By comparing the performance of models pre-trained by different data sizes in Figure 2-Right, we find that the fine-tuning performance of the large models saturates more slowly with the increasing data size compared to the smaller models. For example, the SwinV2-S model pre-trained on IN1K (50%) has a very similar fine-tuning performance to the model pre-trained on IN1K (100%). In comparison, the performance difference between the SwinV2-H model pre-trained on IN1K (50%) and IN1K (100%) is near 0.5, which is a significant gap for ImageNet-1K classification.

Furthermore, a comprehensive observation reveals that the improvements from using more data are not significant under short training lengths. As shown in Figure 5-Right, while there is a noticeable performance gap between SwinV2-H trained on IN1K (50%) and IN1K (100%) at a training length of 500K iterations, the gap is negligible at a training length of 125K iterations. This observation suggests that increasing the training length for larger models is critical to benefit from more data.

Masked image modeling cannot benefit from more data under a non-overfitting scenario. As illustrated in Figure 1-Center, we plot the validation loss of pre-trained SwinV2 models with respect to different dataset sizes, and we demonstrate that for a particular model size and training length, there will be a certain dataset size that will keep it from overfitting. And the model cannot benefit from dataset larger than this size. This behavior differs from previous observations in supervised vision transformers [45] and self-supervised language models [21], where the model performance increases as the number of unique samples increases. We speculate that this may be related to the characteristics of the task itself, i.e., MIM provides a large number of training signals, making the model hard to learn more patterns from large-scale data.

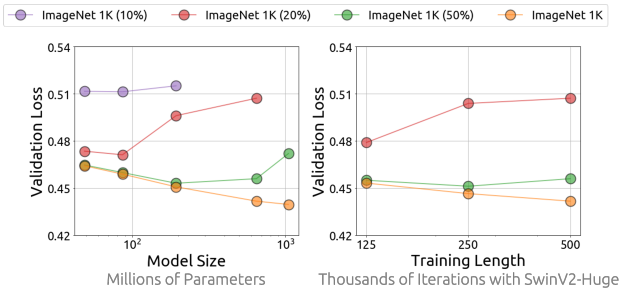

Figure 5. **Left:** Validation loss with respect to model size. The size of the dataset to prevent model from overfitting increases clearly as the model increases. **Right:** Validation loss with respect to training length using SwinV2-Huge. 50% subset of ImageNet-1K is sufficient for training 125K iterations, but not sufficient to prevent overfitting for 500K iterations.*Best viewed in color.*

Evaluation on more tasks. In addition to ImageNet-1K image classification, we also evaluate the MIM pre-trained SwinV2-S, SwinV2-B and SwinV2-L on iNaturalist-18 fine-grained image classification, ADE20K semantic segmentation, and COCO object detection/segmentation. Figure 3 shows a similar pattern with ImageNet-1K (Figure. 2-Right) that as the training cost increases, some models have significantly performance drop. In addition, the smaller models rapidly reach saturation as the amount of data increases, while larger models can continuously benefit from more data after sufficient training. These results suggest that the conclusions drawn on ImageNet-1K are broadly applicable to other vision tasks.

Results with MAE and ViT Figure 4 demonstrates the results of ViT pre-trained with MAE. Similar to SwinV2 models pre-trained with SimMIM, we observe the same

overfitting phenomenon when training with small datasets or large models, which makes MAE still demand for large-scale data. Besides, we could find that larger ViT models can also benefit from more data at a longer training lengths. These experiments verify the methodological and architectural generalizability of our results and findings.

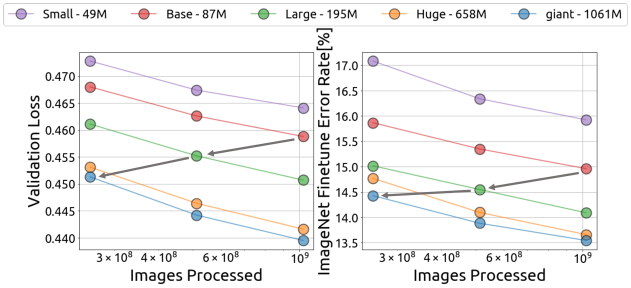

Figure 6. Validation loss and fine-tuning error rate on ImageNet-1K with respect to images processed during pre-training. Large models are more sample efficient than small models to achieve lower validation loss or fine-tuning error rate with fewer optimization steps. *Best viewed in color.*

4.2. Larger Models Possess Higher Sample Efficiency

Figure 6 shows the validation loss and fine-tuning error rate on ImageNet-1K with respect to the images processed (batch size times number of steps) during pre-training stage. Five Swin-V2 models with different sizes are pre-trained on full-set of ImageNet-1K. From these results, we observe that larger models possess higher sample efficiency, reaching the same level of validation loss or fine-tuning error rate on ImageNet with fewer optimization steps. These results indicate that larger models will continue to have better performance, and training larger models with fewer steps on sufficient data is a preferable choice. This observation is also in line with previous works on self-supervised language models [21] and supervised vision transformers [45].

4.3. Correlation between Pre-training Losses and the Fine-tuning Performance

Evaluating a pre-trained model by its fine-tuned performance on downstream tasks is costly. In supervised pre-training, the validation accuracy is used as the proxy indicator to evaluate the quality of the pre-trained models. While in previous studies [7] on other self-supervised learning approaches (*e.g.*, contrastive learning), such a proxy indicator is lacking. In this study, we would like to explore whether the pre-training loss in the training of masked image modeling is a good indicator of its fine-tuning performance. We collect all pre-trained models and plot their training and validation loss curves on Figure 7. Interestingly, the correlations between pre-training losses and the fine-tuning performance on multiple tasks could be observed with a *phase transition* around overfitting.

Task	Overfit		Non-Overfit	
	Train Loss	Val Loss	Train Loss	Val Loss
IN-1K	+0.26	-0.79	-0.64	**-0.90**
iNat-18	+0.17	-0.54	-0.46	**-0.78**
COCO OD	+0.54	-0.81	-0.35	**-0.83**
COCO IS	+0.62	-0.86	-0.31	**-0.85**
ADE-20K	+0.75	-0.91	-0.14	**-0.90**

Table 3. Pearson correlation coefficients between pre-training losses (training and validation losses) and fine-tuning performances on five downstream tasks.

Specifically, the correlation between training loss and fine-tuning performance is negative for the overfitting model (green circles) and positive for the non-overfitting model (red circles). The correlation between validation loss and fine-tuning performance is always negative, but the slope of their linear fit lines [3] is significantly different.

In addition, we further analyze the Pearson correlation coefficient between training loss and fine-tuning performance (Table 3), and find the validation loss has stronger correlation with fine-tuned performance than train loss for all cases, especially for non-overfitting models.

4.4. Effects of Different Sizes of Decoders

We have studied the effects of encoder size from the data scaling perspective. Here, the effects of decoder size are further studied. We pre-train SwinV2-B models with decoder heads of different sizes on IN1K (20%), and Table 4 shows the results. Interestingly, although we find that the heavier decoder has lower training loss and higher validation loss than the linear decoder, indicating a more severe overfitting issue. But there is no decrease in its fine-tuning performance on ImageNet-1K than the linear decoder. This experiment shows that the decoder behaves very differently from the encoder, and we speculate that this is because the decoder "blocks" the damage to the encoder from overfitting.

Decoder	# Params	Train Loss	Val Loss	Top-1 Acc
linear	90.0M	0.46	0.47	84.4
4-blocks	140.4M	0.44	0.48	84.4
8-blocks	190.8M	0.41	0.50	84.5

Table 4. Results of different decoders, including converged training and validation losses of MIM pre-training, and fine-tuning performance (top-1 accuracy) on ImageNet-1K image classification. Encoders for all models are SwinV2-Base.

[3] The least squares method is used for linear fit.

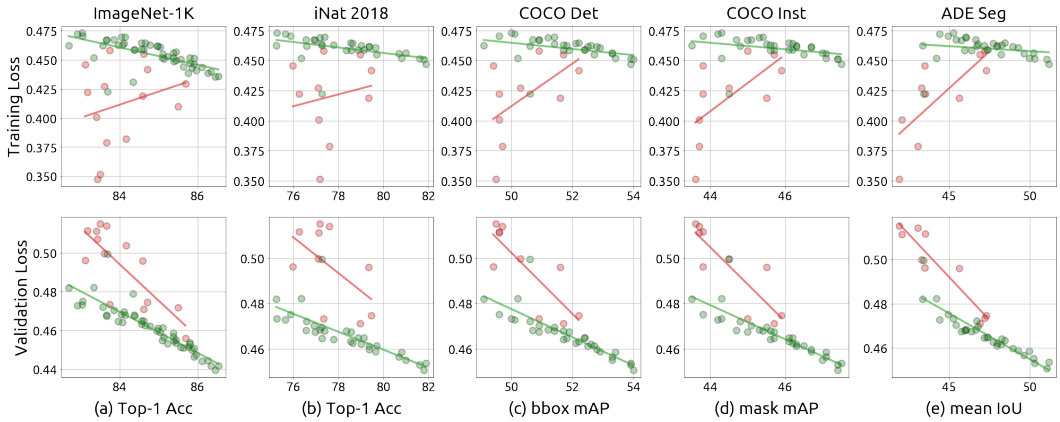

Figure 7. The correlations between pre-training losses (training and validation losses) and the fine-tuning performances of (a) ImageNet-1k image classification, (b) iNat 2018 fine-grained classification, (c) COCO object detection, (d) COCO instance segmentation, (e) and ADE-20K semantic segmentation. Pre-training losses are highly correlated with fine-tuning performance on all five tasks. Red circles indicates the overfitting models and green circles indicates non-overfitting models. *Best viewed in color.*

Dataset	# Classes	Train / Val Loss	Top-1 Acc
IN1K (10%)	1000	0.351 / 0.515	83.5
IN100	100	0.352 / 0.511	83.4

Table 5. Results on different dataset sampling strategies (ImageNet-1K (10%) and ImageNet-100) with same dataset size (1.28×10^5 images for both), include converged training and validation losses of MIM pre-training, and fine-tuning performance (top-1 accuracy) on ImageNet-1K image classification.

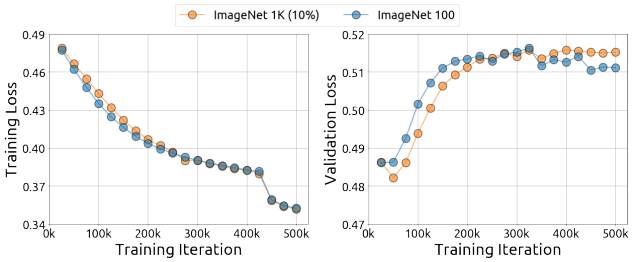

Figure 8. The training loss and validation loss of MIM pre-training with different dataset sampling strategies, ImageNet-1K (10%) and ImageNet-100. *Best viewed in color.*

4.5. Impact of Different Dataset Sampling Strategies

We study different dataset sampling strategies by comparing the training behavior and fine-tuned performance of models pre-trained on IN1K (10%) and IN100. In IN1K (10%), the images are uniformly sampled from each category, and we randomly sample 100 categories from ImageNet-1K as IN100. Experiments are conducted on SwinV2-L with 500K training iterations. Table 5 shows the training loss, validation loss and fine-tuning top-1 accuracy of ImageNet-1K. For the two models pre-trained on IN1K (10%) and IN100, all three metrics are very similar. Figure 8 further illustrates the training dynamics of the two models, and we find both their training loss curves and validation loss curves are almost overlapping. These results show the disparity caused by different dataset sampling strategies is minor.

5. Conclusion

In our work, we systematically study the data scaling capability of masked image modeling at different model sizes and training lengths. Based on the extensive experiments, we demonstrate that in order to scale up computes and model parameters, masked image modeling remains demanding for large-scale data due to the severe overfitting phenomenon when pre-training with small datasets of large models, which challenges the conclusions of previous literatures that a large dataset may not be necessary in masked image modeling. Besides, we also find that masked image modeling cannot benifit from more data under a non-overfitting scenario, which diverges from the previous observations in large-scale pre-training language or vision models and raise new concerns and challenges on the data scalability of MIM. In addition, some intriguing properties of MIM are observed, such as larger models possessing higher sample efficiency and a strong correlation between the validation loss of masked image modeling and the fine-tuning performance. The former observation indicates that larger models will continue to have better performance and suggests that training larger models with fewer steps on sufficient data is a preferable option, while the latter observation suggests that validation loss can be considered as a good proxy indicator for evaluating pre-trained models, and makes it possible to reduce the experimental overhead of measuring models by fine-tuning. We hope that our findings could deepen the understanding of masked image modeling and facilitate future developments on large-scale vision models.

References

[1] Ibrahim Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *NeurIPS*, 2022.

[2] Ibrahim M Alabdulmohsin and Mario Lucic. A near-optimal algorithm for debiasing trained machine learning models. *Advances in Neural Information Processing Systems*, 34:8072–8084, 2021.

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

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[6] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels. *Advances in Neural Information Processing Systems*, 2020.

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

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020.

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[10] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

[11] 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. Ieee, 2009.

[12] 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.

[13] 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 *International Conference on Learning Representations*, 2021.

[14] 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.

[15] 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.

[16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2021.

[17] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

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

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[20] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[22] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

[23] Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on Machine Learning*, pages 5958–5968. PMLR, 2020.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.

[26] 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.

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021.

[28] Microsoft. Turing-nlg: A 17-billion-parameter language model by microsoft, 2020.

[29] Microsoft. Using deepspeed and megatron to train megatron-turing nlg 530b, the world's largest and most powerful generative language model, 2021.

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 2536–2544, 2016.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[34] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *arXiv preprint arXiv:2106.05974*, 2021.

[35] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

[36] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

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

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

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[41] 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.

[42] 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.

[43] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[44] 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. *arXiv preprint arXiv:2111.09886*, 2021.

[45] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2021.

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

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[48] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018.