Efficient MAE towards Large-Scale Vision Transformers

Qiu Han[1], Gongjie Zhang[1,2], Jiaxing Huang[1], Peng Gao[3], Zhang Wei[3], Shijian Lu[1*]

[1]S-Lab, Nanyang Technological University
[2]Black Sesame Technologies
[3]Shanghai Artificial Intelligence Laboratory

han023@e.ntu.edu.sg, Gjz@ieee.org, {Jiaxing.Huang, Shijian.Lu}@ntu.edu.sg, gaopeng@pjlab.org.cn

Abstract

Masked Autoencoder (MAE) has demonstrated superb pre-training efficiency for vision Transformer, thanks to its partial input paradigm and high mask ratio (0.75). However, MAE often suffers from severe performance drop under higher mask ratios, which hinders its potential toward larger-scale vision Transformers. In this work, we identify that the performance drop is largely attributed to the over-dominance of difficult reconstruction targets, as higher mask ratios lead to more sparse visible patches and fewer visual clues for reconstruction. To mitigate this issue, we design Efficient MAE that introduces a novel Difficulty-Flatten Loss and a decoder masking strategy, enabling a higher mask ratio for more efficient pre-training. The Difficulty-Flatten Loss provides balanced supervision on reconstruction targets of different difficulties, mitigating the performance drop under higher mask ratios effectively. Additionally, the decoder masking strategy discards the most difficult reconstruction targets, which further alleviates the optimization difficulty and accelerates the pre-training clearly. Our proposed Efficient MAE introduces 27% and 30% pre-training runtime accelerations for the ViT-Large and ViT-Huge models, provides valuable insights into MAE's optimization, and paves the way for larger-scale vision Transformer pre-training. Code and pre-trained models will be released.

1. Introduction

Vision Transformers [6, 7, 18, 32, 41] have achieved promising performance and emerged as generic models in various computer vision tasks. However, empirical studies [5, 6, 18, 21, 48, 51] reveal that vision Transformers tend to require much more training data than convolutional neural networks (CNNs) [22, 33, 38] due to the lack of inductive bias. Fortunately, recent masked image modeling (MIM) methods [2, 3, 5, 21, 44, 47, 53], which first mask out some patches of the input image and then reconstruct the masked patches, define a promising paradigm in learning representations from unlabeled images and have shown superb performance for vision Transformers.

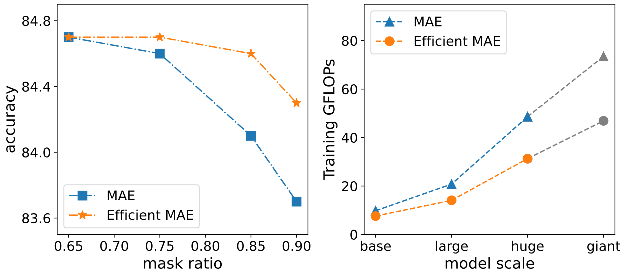

Figure 1. ***Left:*** Comparison of pre-training performance between MAE and our Efficient MAE under different mask ratios. Our proposed Efficient MAE [21] is more robust against higher mask ratios. ***Right:*** Being tolerant to higher mask ratios, Efficient MAE can significantly reduce the computational costs of pre-training, enhancing scalability with larger models and datasets. Results are obtained with ViT-Large on ImageNet-1K. Best viewed in color.

Among these MIM approaches, Masked Autoencoder (MAE) [21] has become prevalent due to its superb pre-training efficiency and good performance. The high pre-training efficiency primarily comes from an asymmetric encoder-decoder pipeline and a high mask ratio (*i.e.*, 0.75), which relieves the heavy encoder from the computation of masked image patches, leading to significant speedups (especially on large models such as 2.8× for ViT-Large and 3.5× for ViT-Huge). The high mask ratio in MAE enables efficient pre-training and plays a key role for scaling up the model size and leveraging large-scale data.

In this work, we investigate how to lift the mask ratio in MAE [21] to further reduce pre-training computational

[*]Corresponding author.

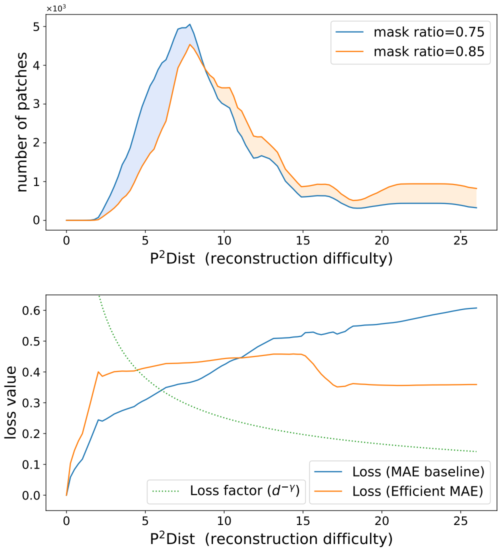

Figure 2. Motivations of our proposed Efficient MAE. P^2Dist is our proposed metric for measuring the reconstruction difficulty. ***Upper:*** MAE pre-training with a higher mask ratio leads to a larger number of hard reconstruction targets. ***Lower:*** Therefore, the pre-training of MAE baseline under higher mask ratios will thus be dominated by the increased hard targets, which hinders the pre-training optimization. To mitigate this issue, our proposed Efficient MAE includes a loss factor $(d^{-\gamma}, \gamma > 0)$ into the MAE baseline, which weights down (or up) the losses for hard (or easy) targets, leading to balanced losses over various reconstruction targets, and facilitating pre-training optimization. Results are obtained with ViT-Large on ImageNet-1K. Best viewed in color.

costs. We start from a simple observation: MAE's pre-training performance drops drastically when the mask ratio grows to a huge value (*e.g.*, 0.90), as shown in Fig. 1 (left). We believe such a performance drop is highly correlated with the optimization difficulty introduced by the extremely high mask ratio. Specifically, as shown in Fig. 2 (upper), an extremely high mask ratio increases the average distances between masked and visible tokens, thus significantly increasing the number of tokens that are difficult to reconstruct. Hence, these challenging reconstruction targets with high uncertainty will dominate the pre-training losses (shown by the blue line in Fig. 2 (lower)), resulting in degraded fine-tuning performance. Based on these observations, we believe that a simple and intuitive solution to mitigate the optimization difficulty under a high mask ratio is to differentiate tokens with different reconstruction difficulties, and balance their losses during pre-training, as illustrated by the orange line in Fig. 2 (lower).

With the motivations above, we propose *Efficient MAE*, which enables effective MAE pre-training under higher mask ratios, and paves the way for more efficient and scalable MIM methods with larger models and datasets. Specifically, we first define a metric named ***P****atch-wise Average Nearest* ***P****ixel* ***Dist****ance* (P^2Dist), which computes the distances between the pixels in the masked patches and the visible pixels, and accordingly reflect the reconstruction difficulties of masked patches. On top of P^2Dist, the proposed Efficient MAE consists of two novel designs. *First*, we design a novel *Difficulty-Flatten Loss*, which mitigates the imbalance between hard and easy targets by down-weighting the loss assigned to abnormally hard targets with high ambiguity and focusing on the pre-training with simple targets, as shown in Fig. 2 (lower). With the proposed Difficulty-Flatten Loss, MAE pre-training becomes less sensitive to the mask ratio (shown in Fig. 1 (left)), pushing the appropriate limit of mask ratio from 0.75 to 0.85. *Second*, we propose a decoder masking strategy to speed up the MAE decoding procedure during pre-training, which selectively reconstructs masked tokens and complements the proposed Difficulty-Flatten Loss in a similar manner. Together with the two designs, our proposed Efficient MAE introduces about 27% and 30% run-time acceleration for ViT-Large and ViT-Huge, respectively.

In summary, the contributions of this work are fourfold.

- We introduce P^2Dist to analyze the optimization of MAE and identify that the pre-training losses over-dominated by those difficult reconstruction targets are the root of performance drop under high mask ratios.
- We propose a novel *Difficulty-Flatten Loss* to down-weigh the pre-training losses for difficult reconstruction targets, thus reducing the ambiguity and reconstruction difficulty in pre-training.
- We propose a decoder masking strategy that selectively reconstructs masked patches, which complements the proposed Difficulty-Flatten Loss and introduces further speed-up to MAE pre-training.
- Based on the two novel designs above, we propose *Efficient MAE*, which achieves about 30 % acceleration on top of the MAE baseline. The proposed Efficient MAE is the pioneering work to investigate MAE pre-training with higher mask ratios for scaling up toward larger models and datasets.

2. Related Work

Large-scale Vision Transformer. Originating from natural language processing (NLP), Transformers have been successfully applied to various computer vision tasks [6, 11, 32, 35, 42, 52, 54], and demonstrated extraordinary potential in scaling model capacity and data size. Numerous visual benchmarks are successively dominated by large-scale Transformers. Early efforts [13, 37, 51] mainly focus on image classification and achieve outstanding performance on ImageNet [15]. The models in these methods contain over a billion parameters and are trained on huge datasets, i.e., JFT-3B. Afterwards, multiple structures of vision Transformer and multimodal data are explored and prevail in more vision tasks, such as object detection [7, 31, 40, 50], semantic segmentation [10, 25, 31, 40], visual question answering [1,43,49,50], etc. The rapidly increasing model capacity and growing sizes of datasets place higher demands on pre-training methods with higher efficiency.

Masked Image Modeling. Inspired by the success of masked language modeling (MLM) (*e.g.*, BERT [16]) in natural language processing (NLP), masked image modeling (MIM) has become a popular trend for vision self-supervised learning. BEIT [5] explores MIM by recovering the masked image into visual tokens from discrete VAE [36]. SimMIM [47], MaskFeat [44], and MAE [21] further demonstrate that low-level visual signals, such as RGB pixel value or the feature descriptor HOG [14], can also be effective reconstruction targets and lead to rich visual representation. Moreover, MAE [21] adopts an asymmetric encoder-decoder framework and partial input scheme, which achieves excellent pre-training efficiency, thus becoming a popular MIM paradigm. After that, the reconstruction target of MAE is widely explored. Data2vec [3], SdAE [9], BootMAE [17], and SIM [39] introduce the momentum encoder to generate the informative reconstruction targets for MIM. In addition, AttMask [24] and SemMAE [26] leverage the attention map or semantic information to guide the masking strategy in MIM, leading to a more effective masked learning process. MC-MAE [20], UM-MAE [27], MixMIM [29], and Green-MIM [23] explore efficient MIM methods for hierarchical ViTs [20, 32, 41]. Different from them, our work focuses on further improving the efficiency of MIM, enabling to pre-train larger-scale visual Transformers in an economic and environmental-friendly way.

Mask Ratio. The mask ratio is a crucial parameter that controls the performance of MIM methods, as well as the efficiency in MAE [21] where an asymmetric encoder-decoder architecture is adopted. Previous studies [19, 21] suggest that mask ratio is highly correlated to the information redundancy of the tasks, thus leading to different mask ratios in different modalities, such as 0.15 in BERT [16] for natural texts, 0.75 in MAE [21] for images, and 0.90 in VideoMAE [19] for videos. Different masking strategies (block-wise and patch-wise masking) and patch sizes also result in different mask ratios (0.40, 0.60, and 0.75 in BEIT [5], SimMIM [47] and MAE [21]) in different MIM methods. Furthermore, DMAE [4] introduces distillation into MAE by aligning the intermediate feature of the student model and that of the pre-trained teacher model. The distillation enables a higher mask ratio than 0.75, but relies heavily on the pre-trained teacher model. ExtreMA [46] utilizes extremely large patch masking (75%-90%) as a strong data augmentation for contrastive learning. Different from previous methods, our work explores the choice of mask ratio from a new perspective – the reconstruction difficulty. By balancing the loss of masked patches with different reconstruction difficulties, we enable MAE pre-training with a higher mask ratio, thus accelerating the MIM pre-training and saving computational resources.

Mask Ratio	0	0.01	0.25	0.50	0.75
Training Gflops	61.6	66.4	51.4	36.0	20.8
Ratio	1	1.07	0.83	0.58	0.34
Memory Usage(G)	38.4	46.0	36.6	27.9	20.7
Ratio	1	1.20	0.94	0.73	0.54

Table 1. GFLOPs and GPU memory usage at different mask ratios. Increasing mask ratio leads to great speedup and memory saving. The model is ViT-large with batch size of 128. Only the encoder is evaluated when mask ratio equals 0. "Ratio" is the rate of computation resources compared to that when mask ratio equals 0.

Mask Ratio	0.75	0.85	0.90
Accuracy	84.6	84.1	83.7
Pre-training Loss	0.407	0.486	0.549

Table 2. Fine-tuning accuracy and pre-training loss under different mask ratios. Higher mask ratios lead to degraded performance and larger loss values. The experiments are conducted with ViT-Large pre-trained for 200 epochs.

3. MAE with Higher Mask Ratios

3.1. Revisiting MAE

MAE [21]. Following ViT [18], MAE [21] first divides each input image into non-overlapping patches and then flattens them into a 1D patch sequence with a length of L. Then, given a mask ratio of r (typically 0.75), $L \cdot r$ patches are randomly masked out, and only the remaining $L \cdot (1-r)$ patches are visible and fed into the ViT encoder. The output tokens of the encoder are concatenated with $L \cdot (1-r)$ learnable $[mask]$ tokens and then processed by a lightweight decoder to reconstruct the masked image patches.

MAE's High Pre-Training Efficiency. MAE [21] is known for high pre-training efficiency, primarily thanks to its high mask ratio. Specifically, only a small portion of visible image patches are used for encoding, which relieves the heavy encoder from the computation of masked visual tokens. As shown in Tab. 1, higher mask ratios can lead to significantly decreased training GFLOPs and GPU memory usage, which enables the pre-training of large and high-capacity models with reduced carbon footprints. In this work, we attempt to increase the mask ratio further to enable efficient pre-training on larger-scale vision Transformers.

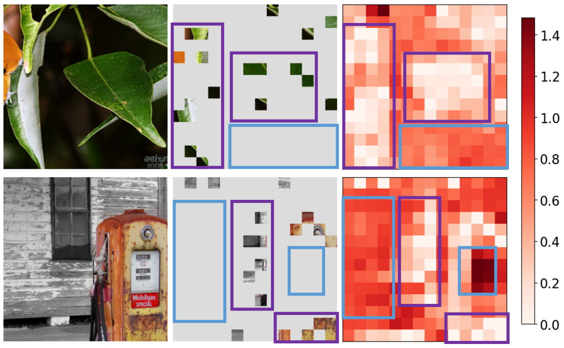

Figure 3. Loss distribution of MAE pre-training under a mask ratio of 0.90. The purple and blue boxes denote areas with sufficient visible patches and vacant areas, respectively. The masked patches tend to have higher reconstruction losses when they have fewer visible patches around them.

3.2. Obstacles with Higher Mask Ratios

Based on the above discussion, one straightforward approach to further advance MAE's efficiency is to raise the mask ratio to a higher value (larger than the default value of 0.75). However, as shown in Tab. 2, raising the mask ratio beyond 0.75 impairs the MAE pre-training performance consistently. The rising losses from the reconstruction task also indicate that higher mask ratios will complicate the reconstruction of corrupted images and further impede the optimization of the vision Transformer encoder.

To better understand the optimization difficulty under high mask ratios, we visualize the spatial distribution of reconstruction loss over the masked images in Fig. 3. It can be observed that, despite the loss distribution being highly related to image contents, those blank regions without any visible patches (blue boxes) usually have higher reconstruction loss compared with regions with visible patches (purple boxes). This phenomenon is intuitive, as it would be simpler for the model to reason the masked patches from the adjacent visible ones, while large blank regions bring substantial uncertainty considering the intricate real-world scenario. We argue that forcing the model to overfit certain specific content without sufficient visual clues is unreasonable and might hinder the optimization of MAE.

The above phenomenon inspires our conjecture that higher mask ratios lead to more challenging targets with high uncertainty, which increases the optimization difficulty and results in a performance drop. Therefore, a quantitative measurement of such reconstruction difficulty and the corresponding scheme to balance the difficulty is crucial for MAE [21] under extremely high mask ratios, which will be detailed in the following subsections.

Algorithm 1 (Approximate) $\mathrm{P^2Dist}$

```
procedure P^2Dist
# $p_c^m$: the center of a masked patch.
# $p_c^v$: centers of the visible patches.
# P: patch size.
Getting S, a set of $k$ nearest visible pixels to $p_c^m$, from :
            $\{p_c^v + \Delta p, \Delta p \in R\}$
    where $R = \{(-\frac{P}{2}, -\frac{P}{2}), (-\frac{P}{2}, 0), ..., (\frac{P}{2}, \frac{P}{2})\}$.
Computing average nearest distance $d'$ between pixels
    in the masked patch and pixels of set S.
return $d'$
```

3.3. Reconstruction Difficulty $\mathrm{P^2Dist}$

As the mask ratio increases, the spatial distribution of visible patches becomes more sparse, leading to greater average distances among visible patches, as well as among masked patches and visible patches. Accordingly, it is intuitive that the reconstruction difficulty of a masked patch should take into account both the density of adjacent visible patches and the distance to the nearest visible patch.

Inspired by SimMIM [47], we design **P**atch-wise Average Nearest **P**ixel **Dist**ance (**P^{2}Dist**) to measure the reconstruction difficulty of different masked patches. $\mathrm{P^2Dist}$ for a masked patch can be formulated as below:

$$d = \frac{1}{P^2} \sum_i^P \sum_j^P \min_{p \in V} (D(p_{i,j}^m, p)) \tag{1}$$

where P indicates the patch size of ViT, V denotes pixels of visible patches, p^m indicates pixel in the masked patch, and $D(\cdot, \cdot)$ is the Euclidean distance function.

However, the computational complexity of $\mathrm{P^2Dist}$ on an image of $H \times W$ is about $O(H^2W^2)$, which is unacceptable. To speed it up, we adopt an approximate distance as shown in Algorithm 1. We only calculate the average nearest distance between pixels in the masked patch and k boundary pixels of nearby visible patches. The overall computation is reduced to $O(kH^2W^2/P^4 + kHW)$. Given the image size of 224 and patch size of 16, the complexity is roughly equivalent to $O(2kHW)$, thus efficient enough for real-time computation. k is set as 8 by default.

We visualize a few masked patches with different $\mathrm{P^2Dist}$ in Fig. 4. The reconstruction difficulties from easy to hard are measured by the rising values of $\mathrm{P^2Dist}$. For easy targets, the masked patch is densely surrounded by visible patches, resulting in small $\mathrm{P^2Dist}$. The masked patch can be easily recovered from the surrounding contents. For difficult targets, $\mathrm{P^2Dist}$ becomes larger as the masked patch is far from visible patches.

Based on reconstruction difficulty metric $\mathrm{P^2Dist}$ discussed above, we visualize the number of masked patches and pre-training losses under different $\mathrm{P^2Dist}$ in Fig. 2 (upper). Two important properties can be observed. *First*, as

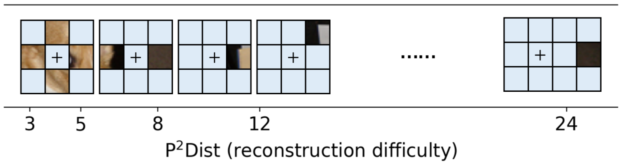

Figure 4. Visualization of masked patches with different P^{2}Dist. The blue blocks denote masked patches. '+' denotes the masked patch which P^{2}Dist is computed on.

the mask ratio increases, patches with lower difficulty levels (P^{2}Dist) decrease, while patches with higher difficulty levels increase significantly. Specifically, by accumulating samples along P^{2}Dist, we find 87% masked patches have at least one visible patch in their 3 × 3 neighborhood when mask ratio is 0.75. However, this percentage drops to 68%/52% with mask ratios of 0.85/0.9. *Second*, predictions of patches with larger P^{2}Dist are more difficult to optimize, leading to larger loss values. These observations are consistent with our analysis in Sec. 3.2, supporting our claim that the difficulty of reconstructing a masked patch is highly correlated with our proposed P^{2}Dist.

4. Efficient MAE

Based on the above findings, we propose **Efficient MAE** that is equipped with a higher mask ratio and accelerated decoder. We first introduce a Difficulty-Flatten loss that alleviates imbalanced reconstruction difficulties and enables pre-training with a higher mask ratio. Besides, we also introduce a decoder masking strategy as a complement to further speed up the decoder.

4.1. Difficulty-Flatten Loss

We propose a Difficulty-Flatten Loss (DFloss) to address the optimization difficulty with higher mask ratios in which there exists a severe imbalance between decreasing easy targets and increasing hard targets. Specifically, we propose to include a modulating factor $d_l^{-\gamma}$ on top of the original reconstruction loss. Taking l_2 loss used in MAE [21] as an example, our proposed DFLoss can be formulated as:

$$L = \frac{1}{\sum_{p_i \in \mathcal{M}} d_i^{-\gamma}} \sum_{p_i \in \mathcal{M}} d_i^{-\gamma} \|p_i - t_i\|_2 \tag{2}$$

where p_i, t_i denote the predicted pixel value and the target value, respectively; $\mathcal{M}$ denotes the set of masked patches; ($\gamma \geq 0$) is the flatting parameter to control the extent to which the hard targets are punished.

As shown in Fig. 2, our proposed DFLoss amplifies losses of easy targets and suppresses losses of hard targets, producing a flatter loss curve (the green line). This flattened loss alleviates the imbalance caused by over-dominated hard targets effectively. By reducing losses of hard targets, it also mitigates the optimization difficulty from the uncertainty of isolated targets with barely any visual clues.

4.2. Decoder Masking

Inspired by [45], we decouple the mask ratio in MAE into two parts: corruption ratio in the encoder and prediction ratio in the decoder. The corruption ratio, which depicts the ratio of masked patches, controls the overall reconstruction difficulty and is discussed in Sec. 3. In this subsection, we mainly discuss the influence of prediction ratio (the portion of masked patches to be predicted), and propose a decoder masking strategy as a complement to DFLoss.

The prediction ratio is considered to affect the optimization of the model [30, 45], since more predictions result in more supervision signals from the loss gradient. However, we find that masked patches with different P^{2}Dist (reconstruction difficulty) contribute unequally to the performance. As our experiments will demonstrate (Sec. 5.3.4), reconstructing easy patches is more effective than reconstructing hard patches. We thus propose a **difficulty-based masking** strategy as a complement to DFLoss, which further alleviates optimization difficulties at high mask ratios. The masked patches are discarded based on a difficulty threshold β before being fed into the decoder. For these extremely hard patches, most of the surrounding information is erased, thus the reconstruction becomes perplexing. Moreover, when equipped with DFLoss, tiny weights will be assigned to such hard targets, making their contribution to the loss much lower (as illustrated by the green line in Fig. 2(lower)). Consequently, discarding patches with P^{2}Dist larger than β has a negligible effect on performance.

5. Experiments

5.1. Experiment Setup

Model Setups. We adopt ViT-Large [18] as the default encoder in ablation studies. Following MAE [21], the decoder consists of 8 Transformer blocks with a dimension of 512. We adopt a patch size of 16 for ViT-Base and ViT-Large, and 14 for ViT-Huge. The default mask ratio is 0.85 if not otherwise stated. The γ in Difficulty-Flatten Loss and β in difficulty-based masking are set to 0.6 and 26 as default.

Training Setups. We follow common pre-training and fine-tuning procedures on ImageNet [15] as previous methods [5, 21, 47], with Top-1 validation accuracy for evaluation. Specifically, models are pre-trained for 200 epochs (short schedule for ablation experiments) or 800 epochs (long schedule for main results) with a batch size of 1024. We adopt AdamW [34] optimizer and cosine learning rate scheduler with a base learning rate of 1.5e-4 and weight decay of 0.05. The learning is linearly increased at the first 40 epochs for warming up. Random resized cropping and

Methods	Partial Input	Backbone	Params.(M)	Mask Ratio	P-Epochs	Min./Epoch	GFLOPs	Acc. (%)
• *Base-size Model*								
SimMIM [47]		Swin-B	88	0.6	800	10.6	11.3	84.0
BEIT [5]		ViT-B/16	88	0.4	800	15.8	18.8(44.9)	83.2
MaskFeat [44]		Vit-B/16	88	0.4	1600	13.5	17.6	84.0
CAE [8]	✓	ViT-B/16	88	0.4	800	13.2	-	83.6
MAE [21]	✓	ViT-B/16	88	0.75	1600	7.5	9.8	83.6
Efficient MAE	✓	ViT-B/16	88	0.85	800	6.0	7.6	83.5
• *Large-size Model*								
SimMIM [47]		Swin-L	197	0.6	800	19.2	26.0	85.4
BEIT [5]		ViT-L/16	304	0.4	800	34.1	63.2(89.3)	85.2
MaskFeat [44]		ViT-L/16	304	0.4	1600	-	61.6	85.7
MAE [21]	✓	ViT-L/16	304	0.75	800	12.1	20.8	85.4
MAE [21]	✓	ViT-L/16	304	0.75	1600	12.1	20.8	85.9
Efficient MAE	✓	ViT-L/16	304	0.85	800	8.9	14.1	85.4
Efficient MAE	✓	ViT-L/16	304	0.85	1600	8.9	14.1	85.7
• *Huge-size Model*								
SimMIM [47]		SwinV2-H	658	0.6	800	-	86.2	85.7
MAE [21]	✓	ViT-H/14	632	0.75	1600	21.1	48.6	86.9
Efficient MAE	✓	ViT-H/14	632	0.85	800	14.9	31.3	86.5

Table 3. Comparisons with existing MIM approaches on ImageNet-1k. “Partial Input” indicates that an asymmetry encoder-decoder structure is adopted and only the visible patches are fed into the encoder. “P-Epochs” indicates the pre-training epochs. “Min./Epoch” denotes the training time (minutes) per epoch. All training times are evaluated on 4x NVIDIA A100 GPUs. For BEIT, the GFLOPs in parentheses also count the DALL-E tokenizer.

random flipping are used for data augmentation. After pre-training, models are fine-tuned for 100/50/50 epochs for ViT-B, ViT-L, and ViT-H. All training time and memory usage are tested on 4 A100 GPUs. GFLOPs are measured with a 224×224 image, except SimMIM [47], which adopts 192×192 input. For MAE [21] and BEIT [5], we adopt their official implementation. Other models are tested based on MMSelfSup [12]. We adjust batch size of each method to fit GPU’s memory constraint and evaluate its training time.

5.2. Main Results

In Tab. 3, we compare the efficiency and accuracy of Efficient MAE with other MIM methods, especially with large-size and huge-size models to demonstrate our advantages in scaling up the model capacity. It can be observed from Tab. 3 that Efficient MAE achieves competitive performances on par with the original MAE [21] with various model capacities. To demonstrate the efficiency of our method, we compare the pre-training time per epoch with different MIM methods. Our proposed Efficient MAE runs significantly faster than other methods, especially for those with extra target generators (i.e., BEIT [5], CAE [8]) or those that process intact input images [44]. Compared to original MAE [21], we achieve 20%, 27%, and 30% acceleration on running time, and reduce training GFLOPs by 23%, 33%, and 36% for ViT-B, ViT-L, and ViT-H. It is noteworthy that our method provides more prominent speedups as the model size becomes larger, which proves its excellent scalability on larger-scale models.

5.3. Ablation Study

5.3.1 Main Components Ablation

To further demonstrate the effectiveness of Efficient MAE, we gradually ablate our components and evaluate their effect on performance and computation resources. As shown by the first two rows of Tab. 4, increasing the mask ratio of original MAE [21] from 0.75 to 0.85 brings superior training efficiency and memory usage but leads to a non-trivial performance drop. Our proposed Difficulty-Flatten Loss and difficulty-based masking strategy mitigate the severe optimization difficulty under high mask ratios, achieve 0.5% gain at a mask ratio of 0.85, and reduces $\sim$30% of both GFLOPs and GPU memory usage, enabling more efficient MIM pre-training. When mask ratio is increased to 0.90, Efficient MAE achieves 84.3%, a competitive performance, by just using approximately half the computational costs of original MAE [21]. It should be noted that our improvement increases with the mask ratio. Under an extreme mask ratio of 0.95, our Efficient MAE outperforms the original MAE by 1.3% accuracy.

5.3.2 Difficulty Measurement

As discussed in Sec. 3.3, the reconstruction difficulty of a masked patch should take into account the density of its surrounding visible patches as well as its distance to the nearest visible patch. Tab. 5 shows the impact of different difficulty metrics on our method. Three extra metrics are studied. The first one, “Number of Patches in 3×3”, only measures

DF	DM	MR	GFLOPs	Mem.(G)	Acc(%)
		0.75	20.8	38.4	84.6
		0.85	14.6	32.8	84.1
✓		0.85	14.2	28.2	84.3
	✓	0.85	14.6	32.8	84.5
✓	✓	0.85	14.2	28.2	84.6
		0.90	11.6	26.9	83.7
✓	✓	0.90	10.3	22.3	84.3
		0.95	8.5	24.6	81.6
✓	✓	0.95	5.9	15.6	82.9

Table 4. Ablation studies on training efficiency and Top-1 accuracy with Difficulty-Flatten Loss (“DF”) and Difficulty-based Decoder Masking (“DM”). “MR” denotes mask ratio. “Mem.” denotes memory usage (G) per GPU. The results without “DF” and “DM” are conducted on original MAE with different mask ratios.

Difficulty Metric	Num. Patches	Acc(%)
Baseline	167	84.1
(a) Number of Patches in 3×3	114	84.3
(b) Sum of Patch Distance in 5×5	159	84.5
(c) Nearest Patch Distance	167	84.4
$\mathbf{P^2Dist}$	150	84.6

Table 5. Comparison of different reconstruction difficulty metrics. “Num. Patches.” indicates the number of patches to be reconstructed in the decoder. (a) denotes the number of visible patches in a 3×3 region around a masked patch. (b) denotes the sum of the inverses of the distances between a masked patch and visible patches in a 5×5 region around the masked patch. (c) denotes the distance between a masked patch and its nearest visible patch.

the density of visible patches in the neighborhood and it assigns zero weight to a masked patch if it has no surrounding patches. This metric improves the baseline MAE with a mask ratio of 0.85 by 0.2%, but is 0.3% lower than our proposed $\mathrm{P^2Dist}$. We conjecture that this metric successfully reduces the overall reconstruction difficulty but fails in evaluating the hard targets. The second metric considers both the distance and density of nearby visible patches, while the range 5×5 is also similar to our difficulty threshold β of 26. However, the easy targets might be over-weighted considering the accumulation in this metric and exponent loss weight in DFLoss, resulting in slightly lower performance. The third metric “Nearest Patch Distance” also fails to differentiate the easy targets (the density of visible patches). Our proposed $\mathrm{P^2Dist}$ outperforms other metrics by better measuring both the easy and hard reconstruction targets.

Method	Acc(%)
Baseline	84.1
(a) Difficulty Thresholding	84.3
(b) Difficulty Scheduling	84.4
(c) Difficulty-based Normalization	84.3
Difficulty-Flatten Loss	84.6

Table 6. Comparison of different schedulers to adjust reconstruction difficulty. (a) The hard targets with $\mathrm{P^2Dist}$ larger than a certain threshold are directly discarded. (b) A curriculum learning schedule is adopted by expanding the range of difficulty and gradually introducing more reconstruction patches from easy to hard during pre-training. (c) the numbers of masked patches in different difficult intervals are counted, and the loss is re-weighted compared to the distribution at a mask ratio of 0.75.

γ	0.0	0.4	0.6	0.8	1.0	2.0
Acc(%)	84.3	84.5	84.6	84.6	84.5	84.3

Table 7. Ablation on parameter γ in Difficulty-Flatten loss.

	β	9	13	**26**	None
(a)	Accuracy(%) *w/o* DFLoss	84.2	84.3	84.3	84.1
(b)	Accuracy(%) *w/* DFLoss	84.2	84.3	84.6	84.5
	Number of Patches	82	114	150	167
	Decoder Pred. Ratio	0.42	0.58	0.75	0.85

Table 8. Ablation on threshold β of $\mathrm{P^2Dist}$ for difficulty-based decoder masking. “None” indicates there is no threshold on difficulty and all masked patches are reconstructed. “Decoder Pred. Ratio” is the prediction ratio of decoder. “DFLoss” denotes Difficulty-Flatten Loss. For “w/o DFLoss”, masked patches are directly discarded based on threshold β, and no modulation is applied to the reconstruction loss.

5.3.3 Difficulty Adjustment

Difficulty Schedulers. In Tab. 6, we study four different schedulers for adjusting reconstruction difficulty. “Difficulty Thresholding” is first adopted so that the reconstruction targets with $\mathrm{P^2Dist}$ higher than a certain threshold are discarded. The performance is slightly improved by only 0.2%, limited by inadequate reconstruction patches. To alleviate this issue, we further investigate “Difficulty Scheduling”, where a curriculum scheduler is applied to gradually increase the task difficulty by expanding the threshold. The performance is further improved by 0.1% but still suffers from the overwhelming hard targets as no additional restriction is imposed. We also evaluate “Difficulty-based Normalization” with a dynamic loss weight as shown in Tab. 6(c). We split the masked patches into different difficulty intervals and normalize the reconstruction loss by comparing the numbers of patches with mask ratios of 0.85 and 0.75 in each interval. Our Difficulty-Flatten Loss outperforms all these schedulers by 0.2% - 0.3% with simpler formation, indicating our method can better balance the supervision from targets with different levels of difficulty.

Decoder Depth	8	4	1
Accuracy(%)	84.6	84.3	84.2

Table 9. Ablation on decoder depth of Efficient MAE.

Loss Parameter γ. Tab. 7 shows the impact of parameter γ in the Difficulty-Flatten Loss. As shown, the fine-tuning performance is robust to a wide range of γ. Note that either a small or a large γ leads to performance drops. When γ equals 0, no regulation is performed on the loss function. While with a large γ, the supervision might focus on extremely easy targets and result in an over-simplified reconstruction task for pre-training.

5.3.4 Decoder Masking Strategy

Difficulty Threshold β. As illustrated in Sec. 4.2, the difficulty threshold β controls the upper limit of reconstruction difficulty in the decoder. Larger β means more masked patches to be reconstructed. Tab. 8 shows impact of varying difficulty threshold β. We first conduct decoder masking without DFLoss as shown by Tab. 8 (a). We observe that the performance hardly changes for different β, indicating that not all patches contribute equally to pre-training. Specifically, those easy targets contribute the majority of the gain. Discarding patches with $\mathrm{P}^2\mathrm{Dist}$ larger than 9 and only reconstructing half of the masked patches achieves 84.2%, surprisingly 0.1% higher than MAE baseline (mask ratio 0.85). Introducing more patches with moderate difficulties slightly improves the performance. However, a small portion of the hardest patches is detrimental to performance, resulting in a 0.2% drop if no threshold is applied.

After applying DFLoss, the impact of $beta$ on the performance changes. When $\beta < 26$, the performance is the same as that without DFLoss. This indicates that DFLoss has little effect on performance with small β, since the difficulties of masked patches are indistinguishable in such cases. However, as β increases, DFLoss enables the model to benefit from reconstructing difficult targets. Raising the threshold from 9 to 26 improves the performance by 0.4%.

Decoder Depth. Tab. 9 shows how the depth of decoder affects our method. Different from MAE [21], a shallow decoder leads to a 0.3% performance drop. We speculate that the appropriate depth of the decoder is highly related to the reconstruction difficulty. The increased mask ratio (from 0.75 to 0.85) leads to more hard patches and severe optimization difficulty. A shallow decoder might hinder the decoder's capacity to reconstruct the hard targets, thereby limiting the improvement of our method which benefits from handling the hard targets. We highlight that our Efficient MAE with just one layer of decoder still surpasses the original MAE under a mask ratio of 0.85.

		ADE20K		COCO	
Methods	Model	mIoU	mAcc	$\mathrm{AP}^{\mathrm{box}}$	$\mathrm{AP}^{\mathrm{mask}}$
MAE	ViT-B	48.1	58.8	50.1	44.7
Efficient MAE	ViT-B	48.0	58.7	50.0	44.5

Table 10. Accuracy on ADE20K semantic segmentation and COCO object detection and segmentation.

5.4. Transfer Learning Experiments

Object detection and segmentation We evaluate the transfer learning capacity of our method on the COCO Dataset following ViTDet [28]. Limited by computation resources, we reran the experiments of the original ViTDet-MAE with a smaller batch size and fewer training epochs (refer to supplementary materials for detailed settings). As shown in Tab. 10, our method achieves comparable results with the original MAE, but with less pre-training cost.

Semantic Segmentation We also experiment on ADE20K using UperNet following MAE. Tab. 10 show that our method performs comparably to the original MAE. Both experiments demonstrate the transfer capability of Efficient MAE on downstream tasks.

6. Conclusion

In this work, we propose Efficient MAE, a new MIM method that is compatible with extremely high mask ratios for efficient pre-training of the large-scale vision Transformer. Our method stems from investigating the performance drop of MAE [21] under higher mask ratios. We identify that the optimization difficulty deriving from the imbalance between easy-hard reconstruction targets is the primary reason for the degradation. To mitigate this issue, we propose a metric $\mathrm{P}^2\mathrm{Dist}$ to measure the reconstruction difficulty of each masked patch. The losses of patches with different $\mathrm{P}^2\mathrm{Dist}$ are thus balanced by our proposed Difficulty-Flatten Loss. Besides, we also propose a decoder masking strategy to discard the hardest targets, further easing the reconstruction difficulty and accelerating the decoder. With the above designs, Efficient MAE significantly accelerates the pre-training of the large model and lessens the burden on computing resources. We hope our study will advance the development of larger visual models, and provide insights into the design of mask ratio for other masked data modeling methods.

Acknowledgement

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 3

[2] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 1

[3] 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. 1, 3

[4] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *arXiv preprint arXiv:2208.12256*, 2022. 3

[5] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image Transformers. In *ICLR*, 2021. 1, 3, 5, 6

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *ECCV*, 2020. 1, 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 1, 3

[8] 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

[9] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. *arXiv preprint arXiv:2208.00449*, 2022. 3

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 3

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2

[12] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021. 6

[13] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 2

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 5

[16] 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. 3

[17] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. *arXiv preprint arXiv:2207.07116*, 2022. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3, 5

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

[20] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 3, 4, 5, 6, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[23] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *arXiv preprint arXiv:2205.13515*, 2022. 3

[24] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv.org*, 2022. 3

[25] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 3

[26] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 3

[27] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022. 3

[28] 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. 8

[29] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual repre-

sentation learning. *arXiv preprint arXiv:2205.13137*, 2022. 3

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5

[31] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *ICCV*, 2021. 1, 2, 3

[33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[35] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 2

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[37] 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. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 2

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[39] 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. 3

[40] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 3

[41] 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 *ICCV*, 2021. 1, 3

[42] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[43] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[44] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 1, 3, 6

[45] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*, 2022. 5

[46] Zhirong Wu, Zihang Lai, Xiao Sun, and Stephen Lin. Extreme masking for learning instance and distributed visual representations. *arXiv preprint arXiv:2206.04667*, 2022. 3

[47] 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. 1, 3, 4, 5, 6

[48] 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

[49] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 3

[50] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[51] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 1, 2

[52] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating DETR convergence via semantic-aligned matching. In *CVPR*, 2022. 2

[53] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 1

[54] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2